Scilight

Complex Systems Stability & Control
https://www.sciltp.com/journals/cssc

Check for updates

Article

Interval Estimation and Attack Detection for CPSs Based on Unknown Input Observer and Watermarking

Zelong Lu, Chaojiang Liang and Zhenhua Wang *

School of Information Science and Engineering, Shandong Normal University, Ji'nan 250014, China
* Correspondence: wzhua111@126.com

How To Cite: Lu, Z.; Liang, C.; Wang, Z. Interval Estimation and Attack Detection for CPSs Based on Unknown Input Observer and Watermarking. *Complex Systems Stability & Control* **2026**, *2*(3), 9. https://doi.org/10.53941/cssc.2026.100020

Received: 28 April 2026
Revised: 28 May 2026
Accepted: 2 June 2026
Published: 30 July 2026

Abstract: This paper proposes an improved proactive defense strategy for secure remote state estimation in cyber-physical systems (CPSs) subject to unknown-but-bounded (UBB) noises, disturbances and man-in-the-middle (MITM) attacks. First, an unknown input observer (UIO) is designed to estimate the system state. To improve the estimation accuracy, a decoupling approach, which can be used to decouple the UBB measurement noise and disturbances, is proposed to overcome the effect of measurement noises and process disturbances in the error system. Then, H_∞ optimization is introduced to enhance the robustness of the error system. Subsequently, an unpredictable and time-varying zonotopic watermarking mechanism is introduced into the sensor-to-estimator communication channel. Furthermore, this proactive defense mechanism operates with strict transparency, eliminating false alarms and preserving nominal estimation performance under attack-free conditions, while inherently forcing stealthy data manipulations to violate predefined detection thresholds. Finally, numerical simulations on a second-order linear system demonstrate that the proposed strategy effectively amplifies malicious interceptions, guaranteeing reliable MITM attack detection against stealthy MITM attacks.

Keywords: cyber-physical systems (CPSs); man-in-the-middle (MITM) attacks; state estimation; unknown-but-bounded (UBB) noises; watermarking

1. Introduction

In recent years, cyber-physical systems (CPSs), which deeply integrate computational resources with physical processes, have been widely deployed in safety-critical infrastructures such as power systems and aerospace [1]. However, the reliance on open communication networks makes CPSs highly susceptible to malicious cyber attacks. Among these vulnerabilities, man-in-the-middle (MITM) attacks pose a severe risk, as adversaries can stealthily intercept and manipulate data transmission channels to launch remote attacks [2–5]. Consequently, secure state estimation is necessary for preserving the operational reliability of CPSs in the face of such adversarial interventions [6–11]. While traditional observers often rely on the assumption of Gaussian noise, practical industrial environments are predominantly subject to unknown-but-bounded (UBB), disturbances and measurement noises [12–15].

To address these challenges, interval state estimation has emerged as an effective framework for providing guaranteed state bounds and enhancing robustness under UBB uncertainties [16–20]. Interval estimation has attracted significant attention for bounding system states under the UBB measurement noise and disturbances [21]. These approaches are typically categorized based on various set representations, such as polytopes, ellipsoids, zonotopes, and intervals. As pointed out in [22], zonotopes provided superior accuracy and more compact representations than ellipsoidal sets, while incurring lower computational complexity compared to polytopes. Consequently, unknown input observers (UIO) have been widely deployed to decouple disturbances and UBB noises. For instance, a novel interval estimation method integrating a less conservative UIO with reachability analysis was proposed in [23] to effectively decouple unknown disturbances, thereby achieving state bounds with enhanced precision and reduced computational complexity. Similarly, the work [24] achieved observability via measurement ensemble and mitigated sensor faults via redundancy using a distributed UIO network. However, estimation accuracy remains unaddressed in the aforementioned literature. If an adversary meticulously designs malicious inputs within the predefined threshold bounds, such as implementing MITM attacks [25–27], traditional interval estimation methods will be severely compromised, leading to undetected system degradation.

To counter these threats, substantial research has been dedicated to advanced attack detection mechanisms [28,29]. In [30], a T-S fuzzy-based H_∞ observer was proposed for attack detection and reconstruction in autonomous underwater vehicles subject to UBB disturbances and measurement noise. Notably, an improved UIO equipped with a disturbance-decoupling strategy and reachability analysis was developed in [31] to facilitate reduced-complexity, residual-based attack detection and reconstruction for CPSs. However, as indicated in [32], such conventional MITM attack detection methods primarily rely on passive residual-based detectors or set-theoretic threshold analysis. As noted in [33], the fundamental distinction between an ordinary fault and a stealthy cyber attack is that the latter is deliberately engineered to evade these passive detection bounds. Consequently, there is a critical need to transition from passive monitoring to proactive defense mechanisms—such as dynamic watermarking, which has proven highly effective for authenticating signals and ensuring data integrity in networked control systems [34–36] to achieve superior detection sensitivity without degrading nominal performance. By injecting a private, unpredictable and time-varying signal into the communication channel, defenders can actively excite the system [37]. If an attacker compromises the watermarked data without knowing the secret pattern, the statistical properties or the bounded nature of the recovered residual will inevitably be distorted, thereby exposing the attack. Recently, researchers have begun integrating set-membership theories with active defenses, exploring zonotopic watermarking to detect replay and MITM attacks. Despite these advancements, it remains an open challenge to design a strictly transparent watermarking mechanism for CPSs operating under UBB noises—one that guarantees reliable detection of stealthy MITM attacks without compromising the nominal estimation performance or increasing the false alarm rate.

Different from the attack-oriented analysis in [4] and the passive residual-based detection framework in [32] this paper focuses on a defender-side proactive detection strategy. Moreover, compared with the watermark-based proactive defense method in [34], the proposed approach does not treat watermarking as an independent add-on mechanism. Instead, the watermark is redesigned on the basis of the UIO-based interval estimator. Specifically, the UIO first decouples the UBB process disturbance and measurement noise and generates tighter residual bounds through zonotopic propagation. Then, the watermark parameters are designed according to these tightened bounds, so that the injected watermark is strictly transparent under attack-free conditions while forcing stealthy MITM manipulations to exceed the precomputed detection thresholds.

Motivated by the above discussions, this paper proposes an improved proactive attack defense strategy to address the secure estimation problem. By integrating a UIO with an improved dynamic watermark approach, this work aims to guarantee reliable detection of MITM attacks in CPSs environments constrained by UBB noises and disturbances. Zero-value injection attacks, requiring merely signal nullification capability, represent a worst-case baseline for evaluating detectability limits of proactive defenses, as they can bypass traditional residual-based detectors when nominal bounds are conservatively wide [4,32]. The main contributions of this article are summarized as follows:

(1) A decoupling approach based on UIO is proposed to eliminate the effect of UBB noises and disturbances from the error system. The proposed approach relaxes the design conditions of the decoupling matrix. Furthermore, integrated with H_∞ optimization, it achieves higher state estimation accuracy, ensuring that the true system states are strictly bounded within dynamically propagated zonotopes.

(2) A time-varying, unpredictable zonotopic watermarking mechanism is designed for proactive attack defense. It is strictly transparent (zero false alarm) under attack-free conditions, while effectively amplifying stealthy MITM and replay attacks to violate precomputed thresholds. Numerical simulations demonstrate that for zero-value injection attacks, the detection rate increases from 0% (watermark-free baseline) to guaranteed reliable detection.

The remainder of this paper is organized as follows. Section 2 introduces the preliminaries and problem formulation. Section 3 presents the main results in this paper. Numerical simulations are provided in Section 4.

Notation 1. *$\mathbb{R}^n$ represents a vector with n dimensions and $\mathbb{R}^{m\times n}$ represents a matrix with m rows and n columns. The set of positive integers is $\mathbb{Z}^+$. I_n stands for $n \times n$ identity matrix and $\emptyset$ stands for the empty set. The transpose of a matrix A is represented as $A^\top$ and the pseudo-inverse of a matrix A is symbolized by $A^\dagger$. $P \succ 0$ ($P \prec 0$) denotes that the positive (negative) definite matrix. Within a symmetric partitioned matrix, the asterisk $*$ is used to represent the transpose of a matrix that is symmetric with respect to its main diagonal. $\|A\|_2$ is denoted as the 2-norm of the matrix A. $\mathrm{supp}(A)$ represents the support function of A.*

2. Preliminaries and System Description

Definition 1 ([37]). *An interval vector $\mathbb{I} \in \mathbb{R}^n$, also known as a box, is represented as*

$$\mathbb{I} = [p, q] = \{x : x \in \mathbb{R}^n, p_i \leq x_i \leq q_i\},$$

where $p_i = [p_1, \ldots, p_n]^T$ and $q_i = [q_1, \ldots, q_n]^T$.

Definition 2 ([37]). *An i-order zonotope $\mathbb{Z} \subseteq \mathbb{R}^n$ is defined as*

$$\mathbb{Z} = \langle p_z, H_z \rangle = \left\langle p_z + H_z z : z \in \mathbf{B}^i \right\rangle,$$

where $\mathbb{Z}$ is the affine transformation of the hypercube $\mathbf{B}^i = [-1, 1]^i$, $p_z \in \mathbb{R}^n$ is the center of the zonotope $\mathbb{Z}$ and $H_z \in \mathbb{R}^{n \times m}$ is known as the generating matrix of the zonotope $\mathbb{Z}$.

Definition 3 ([34]). *For a set $\mathbf{G} \subseteq \mathbb{R}^n$, its interval hull is defined as the smallest interval vector enclosing it, which is represented as*

$$\mathbf{G} \subseteq \mathrm{Box}(\mathbf{G}) = [c, d],$$

where $c = [c_1, \ldots, c_n]^T$ and $d = [d_1, \ldots, d_n]^T$.

Definition 4 ([38]). *The Minkowski sum of two sets $\mathbf{Y}$ and $\mathbf{Z}$ is defined as*

$$\mathbf{Y} \oplus \mathbf{Z} = \{x + z : x \in \mathbf{Y}, z \in \mathbf{Z}\},$$

where $\oplus$ denotes the Minkowski sum.

Property 1 ([38]). *For zonotopes, the following equations can be obtained as*

$$\mathbb{Z} = \mathbb{Z}_i \oplus \mathbb{Z}_j = (p_i + p_j) \oplus \begin{bmatrix} H_i & H_j \end{bmatrix} \mathbf{B}^{i+j},$$

$$L_1 \mathbb{Z} = L_1 \langle p, H \rangle = L_1 \odot \langle p, H \rangle = \langle L_1 p, L_1 H \rangle.$$

Property 2 ([39]). *For a series of sets $S_i \subset \mathbb{R}^n$, $i = 1, \ldots, m$, the following equation holds*

$$\mathrm{Box}\left(\bigoplus_{i=1}^{m} S_i \right) = \bigoplus_{i=1}^{m} \mathrm{Box}(S_i).$$

In this work, the following discrete-time linear system with UBB process disturbances and measurement noises is considered:

$$\begin{cases} x_{k+1} = A x_k + B u_k + E w_k, \\ y_k = C x_k + F v_k, \end{cases} \tag{1}$$

where $x_k \in \mathbb{R}^{n_x}$, $u_k \in \mathbb{R}^{n_u}$ and $y_k \in \mathbb{R}^{n_y}$ denote the vectors of state,control input and measurement output, respectively. A, B, E, C and F are system matrices with appropriate dimensions. $w_k \in \mathbb{R}^{n_w}$ and $v_k \in \mathbb{R}^{n_v}$ denote the unknown disturbances and measurement noise, respectively. Without loss of generality, x_0, w_k and v_k are assumed to be unknown but bounded as

$$\begin{aligned} x_0 &\in \mathcal{X}_0 = \langle p_0, H_0 \rangle, \\ w_k &\in \mathcal{W} = \langle 0, H_w \rangle, \\ v_k &\in \mathcal{V} = \langle 0, H_v \rangle, \end{aligned}$$

where $p_0 \in \mathbb{R}^{n_x}$ and $H_0 \in \mathbb{R}^{n_x \times n_x}$, $H_w \in \mathbb{R}^{n_w \times n_w}$, $H_v \in \mathbb{R}^{n_v \times n_v}$ are known matrices and vectors. The unknown-but-bounded (UBB) assumption is widely adopted in set-membership and interval observer design, since in many practical engineering systems the exact statistical distributions of disturbances and measurement noises are unavailable, whereas their admissible amplitude ranges can often be estimated from physical constraints, sensor specifications, or historical data [40–42]. The objective is to estimate the upper and lower bounds of x_k, i.e. to find an interval vector $[\underline{x}_k, \overline{x}_k]$ such that $\underline{x}_k \leq x_k \leq \overline{x}_k$.

Lemma 1 ([43])**.** *The linear matrix inequality satisfies the following condition:*

$$\begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix} \succ 0, \tag{2}$$

where X and Y are symmetric matrices, that is, $Y = Y^\top$, $X = X^\top$, then (2) is equivalent to

$$Y \succ 0, \quad X - ZY^{-1}Z^\top \succ 0. \tag{3}$$

Lemma 2 ([44])**.** *For matrices $X \in \mathbb{R}^{a\times b}$, $Y \in \mathbb{R}^{b\times c}$, and $Z \in \mathbb{R}^{a\times c}$, if $\operatorname{rank}(Y) = c$, the minimum-norm solution of the matrix equation $XY = Z$ is obtained as*

$$X = ZY^\dagger, \tag{4}$$

where $Y^\dagger \in \mathbb{R}^{c\times b}$ is the Moore-Penrose pseudoinverse of Y.

3. Main Results

3.1. The Design of UIO Architecture

In this section, a UIO is constructed at the sensor to decouple both the unknown process disturbance w_k and the measurement noise v_k for robust state estimation. The observer generates the state estimate $\hat{x}_k$ using only the local input u_k and measurement y_k. Motivated by the classical unknown input observer framework in [45–47], the following state estimator is adopted:

$$\hat{x}_k = TA\hat{x}_{k-1} + TBu_{k-1} + L(y_{k-1} - C\hat{x}_{k-1}) + Ny_k, \tag{5}$$

where $\hat{x}_k \in \mathbb{R}^{n_x}$ is the estimation of x_k, $L \in \mathbb{R}^{n_x\times n_y}$ is the observer gain, and $T \in \mathbb{R}^{n_x\times n_x}$, $N \in \mathbb{R}^{n_x\times n_y}$ are constant matrices designed to satisfy

$$T + NC = I_{n_x}. \tag{6}$$

Substituting $T = I_{n_x} - NC$ from (6), the condition $TE = O$ is imposed to decouple the process disturbance w_k, while $NF = O$ eliminates the instantaneous output noise. That is,

$$TE = O_{n_x\times n_w}, \qquad NF = O_{n_x\times n_v}. \tag{7}$$

From $T = I - NC$ and (7), it follows that $(I - NC)E = O$, i.e., $NCE = E$. Combining this with $NF = O$ yields the single linear matrix equation for N:

$$N\begin{bmatrix} CE & F \end{bmatrix} = \begin{bmatrix} E & O_{n_x\times n_v} \end{bmatrix}. \tag{8}$$

Remark 1. *It should be noted that (6) alone can always be satisfied for any output matrix C, for example, by choosing $N = 0$ and $T = I_{n_x}$. However, in the proposed UIO design, T and N are required to satisfy (6) together with the decoupling conditions $TE = 0$ and $NF = 0$. Therefore, the existence of such matrices is not guaranteed for an arbitrary output matrix C.*

From $T = I_{n_x} - NC$, the conditions $TE = 0$ and $NF = 0$ can be equivalently written as

$$N[CE \ \ F] = [E \ \ 0].$$

By defining $Q = [CE \ \ F]$ and $S = [E \ \ 0]$, the above equation becomes $NQ = S$. According to Lemma 2, a solution N exists if and only if

$$\operatorname{rank}(Q) = \operatorname{rank}\begin{bmatrix} Q \\ S \end{bmatrix}.$$

When this rank condition holds, a minimum-norm solution can be obtained as

$$N = SQ^\dagger, \tag{9}$$

where $Q^\dagger$ denotes the Moore-Penrose pseudoinverse of Q.

More generally, all feasible solutions can be parameterized as $N = SQ^\dagger + S_0\Pi$, where $\Pi = I_{n_y} - QQ^\dagger$ and

$S_0 \in \mathbb{R}^{n_x \times n_y}$ is a free matrix. Then, T is calculated from $T = I_{n_x} - NC$.

To obtain the error dynamics, the true state is rewritten using $T + NC = I$ and the system Equation (1) as

$$\begin{aligned} x_k &= (T + NC)x_k = Tx_k + NCx_k \\ &= T(Ax_{k-1} + Bu_{k-1} + Ew_{k-1}) + N(y_k - Fv_k) \\ &= TAx_{k-1} + TBu_{k-1} + TEw_{k-1} \\ &\quad + Ny_k - NFv_k. \end{aligned} \tag{10}$$

Define the estimation error as $e_k = \hat{x}_k - x_k$. It follows that

$$e_k = x_k - \hat{x}_k. \tag{11}$$

Subtracting (10) from (5) yields the error dynamics

$$\begin{aligned} e_k &= (TA - LC)e_{k-1} + TEw_{k-1} - LFv_{k-1} - NFv_k \\ &= (TA - LC)e_{k-1} - LFv_{k-1}, \end{aligned} \tag{12}$$

where the term TEw_{k-1} vanishes due to (7) and $NF = 0$ is used.

If the bounds of e_k, i.e., $\underline{e}_k$ and $\bar{e}_k$, can be estimated such that $\underline{e}_k \leq e_k \leq \bar{e}_k$, then from (11), the interval estimation of x_k can be obtained by

$$\begin{cases} \bar{x}_k = \hat{x}_k + \bar{e}_k, \\ \underline{x}_k = \hat{x}_k + \underline{e}_k. \end{cases} \tag{13}$$

Define the residual of the state estimation as follows:

$$r_k = y_k - C\hat{x}_k, \tag{14}$$

which is sent to the remote estimator via a wireless network that is subject to cyber attacks.

While the decoupling conditions guarantee elimination of the disturbance and instantaneous noise, the term $-LFv_{k-1}$ indicates that delayed measurement noise persists in the error dynamics. An H_∞-optimal observer is then derived via a convex LMI to suppress this residual noise channel. The following theorem provides the explicit LMI condition and the resulting observer gain.

Theorem 1. *Given a scalar $\gamma > 0$, if there exists a positive definite matrix $P \in \mathbb{R}^{n \times n}$ ($P \succ 0$), a matrix $W \in \mathbb{R}^{n \times p}$, and a matrix $Y \in \mathbb{R}^{n \times p}$ satisfying the following linear matrix inequality:*

$$\begin{bmatrix} I_n - P & 0 & \xi_1^\top \\ 0 & -\gamma^2 I_p & \xi_2^\top \\ \xi_1 & \xi_2 & -P \end{bmatrix} \prec 0, \tag{15}$$

where

$$\begin{aligned} \xi_1 &= PA - PSQ^\dagger CA - Y\Pi CA - WC, \\ \xi_2 &= -WF, \end{aligned}$$

then the UIO (5) serves as a robust observer for system (1) and the corresponding error system is asymptotically stable and satisfies the H_∞ performance with attenuation level γ, i.e.,

$$\sum_{k=0}^{\infty} \|e_k\|_2^2 < \gamma^2 \sum_{k=0}^{\infty} \|v_k\|_2^2$$

under zero initial condition.

Moreover, the matrices T, N, and L can be obtained by:

$$\begin{aligned} N &= SQ^\dagger + S_0, \\ T &= I_{n_x} - NC, \\ L &= P^{-1}W, \end{aligned} \tag{16}$$

where $S_0 = P^{-1}Y$, $Q = [CE \;\; F]$, $S = [E \;\; 0]$, and $\Pi = I_{n_y} - QQ^\dagger$.

Proof. Consider the Lyapunov function

$$V(k) = e_k^\top P e_k, \quad P \succ 0. \tag{17}$$

To guarantee the H_∞ performance of the error dynamics, the following performance index is introduced:

$$J = \sum_{k=0}^{\infty} \left(\Delta V(k) + e_k^\top e_k - \gamma^2 v_k^\top v_k \right), \tag{18}$$

where $\Delta V(k) \triangleq V(k+1) - V(k)$. A sufficient condition to ensure $J < 0$ for all nonzero v_k is given by

$$\Delta V(k) + e_k^\top e_k - \gamma^2 v_k^\top v_k < 0. \tag{19}$$

Substituting (17) into (19) yields

$$e_{k+1}^\top P e_{k+1} - e_k^\top P e_k + e_k^\top e_k - \gamma^2 v_k^\top v_k < 0. \tag{20}$$

Recall the error dynamics from (12): $e_{k+1} = \Phi e_k - LF v_k$, where $\Phi = (TA - LC)$. Substituting into (20) gives

$$(\Phi e_k - LF v_k)^\top P (\Phi e_k - LF v_k) - e_k^\top P e_k + e_k^\top e_k - \gamma^2 v_k^\top v_k < 0.$$

Expanding the quadratic form, we obtain

$$\begin{aligned} & e_k^\top (\Phi^\top P \Phi - P + I) e_k \\ & \quad - 2 e_k^\top \Phi^\top P L F v_k \\ & \quad + v_k^\top \left((LF)^\top P L F - \gamma^2 I \right) v_k < 0. \end{aligned}$$

This inequality is equivalent to the following matrix inequality:

$$\begin{bmatrix} \Phi^\top P \Phi - P + I & -\Phi^\top P L F \\ -(LF)^\top P \Phi & (LF)^\top P L F - \gamma^2 I \end{bmatrix} \prec 0. \tag{21}$$

Applying the Schur complement lemma to (21), since $P > 0$, an equivalent LMI is obtained:

$$\begin{bmatrix} I_n - P & 0 & \Phi^\top P \\ 0 & -\gamma^2 I & -(LF)^\top P \\ P\Phi & -PLF & -P \end{bmatrix} \prec 0. \tag{22}$$

With the change of variables $W = PL$, it follows that $PLF = WF$. From the definition of Φ, it follows that $P\Phi = P(TA - LC) = PTA - WC$.

Substituting the explicit expression for T from (16) into PTA yields:

$$\begin{aligned} PTA &= P[I_{n_x} - (SQ^\dagger + S_0 \Pi) C] A \\ &= PA - PSQ^\dagger CA - PS_0 \Pi CA. \end{aligned}$$

Since $Y = PS_0$, one has

$$PTA = PA - PSQ^\dagger CA - Y \Pi CA.$$

By defining $Y = PS_0$ (which corresponds to $S_0 = P^\dagger Y$ since $P > 0$), the explicit expression for $P\Phi$ is obtained as:

$$P\Phi = PA - PE\Theta^\dagger CA - Y\Psi CA - WC \triangleq \xi_1.$$

Similarly, define $\xi_3 = -PLF = -WF$. Substituting ξ_1 and ξ_3 into (22) directly yields the LMI condition (15). Therefore, if (15) holds, then (21) holds, which implies (19) and consequently $J < 0$. Summing (19) from $k = 0$ to T_f gives

$$V(T_f + 1) - V(0) + \sum_{k=0}^{T_f} \|e_k\|_2^2 - \gamma^2 \sum_{k=0} \|v_k\|_2^2 < 0.$$

Since $V(T_f+1) \geq 0$, it follows that

$$\sum_{k=0}^{T_f} \|e_k\|_2^2 < \gamma^2 \sum_{k=0}^{T_f} \|v_k\|_2^2 + V(0).$$

Letting $T_f \to \infty$, the H_∞ energy performance is obtained. In particular, under zero initial condition,

$$\sum_{k=0}^{\infty} \|e_k\|_2^2 < \gamma^2 \sum_{k=0}^{\infty} \|v_k\|_2^2.$$

The matrices T and N given in (16) are obtained by solving the linear Equations (6) and (7) using Lemma 2, and the observer gain L is recovered from W as $L = P^{-1}W$. This completes the proof. □

Remark 2. *The H_∞ performance of the UIO can be optimized by solving the following convex optimization problem:*

$$\begin{array}{ll} \min & \gamma^2 \\ \textit{s.t.} & (15). \end{array} \tag{23}$$

The optimal γ provides the minimum disturbance attenuation level. The decision variables are the symmetric positive-definite matrix P, and matrices W and Y, which parameterize the observer gain $L = P^{-1}W$ and the matrix $S_0 = P^{\dagger}Y$, respectively.

3.2. Dynamic Zonotope Watermarking

In this section, a proactive watermarking mechanism is integrated into the data transmission process to actively authenticate the signal integrity. Unlike conventional approaches that assume stochastic noise with known distributions, the proposed watermark is specifically designed for the framework of UBB process disturbances and measurement noises and represented as a zonotope. The watermark is time-varying and kept secret from the adversary, thereby forcing any tampering to manifest as an abnormal deviation in the recovered residual.

At each time step k, a watermark signal $\delta_k \in \mathbb{R}^{n_y}$ is generated by a cryptographically secure pseudorandom number generator (PRNG) and is strictly confined to a known, time-varying zonotope [34–37]:

$$\delta_k \in \mathcal{W}_k^{\Delta} = \langle p_k^{\Delta}, H_k^{\Delta} \rangle \subseteq \text{Box}(\mathcal{W}_k^{\Delta}), \tag{24}$$

where $p_k^{\Delta} \in \mathbb{R}^{n_y}$ is the center vector, $H_k^{\Delta} \in \mathbb{R}^{n_y \times n_g}$ is the generator matrix with n_g denoting the number of generators, and $\text{Box}(\cdot)$ denotes the interval hull operation [37].

The sequence $\{\delta_k\}$ is generated as follows: at each instant, the PRNG produces a random vector $z_k \in [-1,1]^{n_g}$ uniformly distributed over the unit hypercube, and the watermark is obtained via the affine transformation:

$$\delta_k = p_k^{\Delta} + H_k^{\Delta} z_k. \tag{25}$$

It is assumed that the legitimate sensor and remote estimator share a secret seed, which is unavailable to the adversary. The watermark sequence is generated by a cryptographically secure PRNG using this shared secret and the common time index (or packet sequence number). Hence, both legitimate sides can independently reproduce the same δ_k without transmitting the watermark itself. In practice, the shared seed can be established during system initialization through an authenticated secure channel or a standard key-management protocol, while synchronization can be maintained by time stamps or packet counters.

The smart sensor computes the original residual $r_k = y_k - C\hat{x}_k$ and modifies it by adding the watermark before transmission:

$$\tilde{r}_k = r_k + \delta_k. \tag{26}$$

Upon reception, the remote estimator recovers the residual by subtracting the identical locally-regenerated watermark:

$$\hat{r}_k = \tilde{r}_k^{\text{rcv}} - \delta_k, \tag{27}$$

where $\tilde{r}_k^{\text{rcv}}$ denotes the received signal, which may be subject to malicious manipulations [4,5]. Under attack-free conditions, $\tilde{r}_k^{\text{rcv}} = \tilde{r}_k$, leading to $\hat{r}_k = r_k$, which guarantees the strict transparency of the defense strategy [34–36].

Remark 3. *It is worth noting that under attack-free conditions, the received signal is exactly* $\tilde{r}_k^{rcv} = \tilde{r}_k = r_k + \delta_k$*. Subtracting the locally generated watermark* δ_k *yields* $\hat{r}_k = r_k$ *for all k. This inherent cancellation guarantees that the dynamic watermarking mechanism is strictly transparent under nominal operations. Consequently, it affects neither the state estimation performance nor the detection thresholds designed from watermark-free data, thereby ensuring that the false alarm rate remains exactly zero.*

To mathematically guarantee that stealthy data manipulations are deterministically detected, the watermark parameters must be optimally designed, as formalized in the following theorem.

Theorem 2. *Consider the CPSs (1) equipped with the UIO (5) and the residual-based detector with steady-state analytical thresholds* $[\underline{r}_i, \overline{r}_i]$ *for the* i*-th channel (*$i = 1, \ldots, n_y$*). Suppose the adversary launches a zero-value injection MITM attack, such that the received signal is forced to zero (*$\tilde{r}_k^{rcv} = 0$*). The stealthy attack is deterministically detected with a* 100% *detection rate if the watermark parameters* p_k^{Δ} *and* H_k^{Δ} *satisfy the following condition for each channel* i*:*

$$-p_{k,i}^{\Delta} - \sum_{j=1}^{n_g} |(H_k^{\Delta})_{ij}| > \overline{r}_i, \tag{28}$$

or

$$-p_{k,i}^{\Delta} + \sum_{j=1}^{n_g} |(H_k^{\Delta})_{ij}| < \underline{r}_i, \tag{29}$$

where $p_{k,i}^{\Delta}$ *denotes the* i*-th component of* p_k^{Δ}*, and* $(H_k^{\Delta})_{ij}$ *denotes the* (i,j)*-th element of* H_k^{Δ}*.*

Proof. Under a zero-value injection attack, the adversary intercepts the communication channel and replaces the transmitted signal with zero [4]. Consequently, according to the recovery mechanism in (27), the recovered residual evaluated by the remote detector is:

$$\hat{r}_k = 0 - \delta_k = -\delta_k. \tag{30}$$

Since the watermark signal is bounded by the zonotope $\delta_k \in \langle p_k^{\Delta}, H_k^{\Delta} \rangle$, the recovered residual is inherently confined to the following set:

$$\hat{r}_k \in \langle -p_k^{\Delta}, -H_k^{\Delta} \rangle. \tag{31}$$

By calculating the interval hull of this zonotope using Definition 4, the components of the recovered residual for the i-th channel satisfy $\hat{r}_{k,i} \in [\zeta_{k,i}^1, \zeta_{k,i}^2]$, where the exact boundary bounds are derived as:

$$\zeta_{k,i}^1 = -p_{k,i}^{\Delta} - \sum_{j=1}^{n_g} |(H_k^{\Delta})_{ij}|, \tag{32}$$

$$\zeta_{k,i}^2 = -p_{k,i}^{\Delta} + \sum_{j=1}^{n_g} |(H_k^{\Delta})_{ij}|. \tag{33}$$

To ensure reliable detection, the recovered residual interval must strictly fall outside the nominal analytical thresholds $[\underline{r}_i, \overline{r}_i]$. This requires the entire interval $[\zeta_{k,i}^1, \zeta_{k,i}^2]$ to have an empty intersection with the threshold bounds, which is equivalent to:

$$\zeta_{k,i}^1 > \overline{r}_i \quad \text{or} \quad \zeta_{k,i}^2 < \underline{r}_i. \tag{34}$$

Substituting (32) and (33) into (34) directly yields the optimal design conditions (28) and (29). The proof is completed. □

Remark 4. *Although zero-centered watermark generators are often used in stochastic watermarking schemes, deterministic detection under the interval threshold requires the recovered residual interval to be completely separated from the nominal threshold. Therefore, in the proposed deterministic design, the watermark center* p_k^{δ} *and generator radius should be selected according to (28) or (29), rather than simply choosing a zero-centered watermark.*

Remark 5. *The practical selection of the watermark amplitude should balance detection sensitivity and implementation cost. Denote*

$$g_{k,i} = \sum_{j=1}^{n_g} |(H_k^\delta)_{ij}|,$$

as the radius of the watermark interval in the i-th channel. Under a zero-value injection attack, the recovered residual satisfies

$$\hat{r}_{k,i} \in [-p_{k,i}^\delta - g_{k,i}, -p_{k,i}^\delta + g_{k,i}].$$

Therefore, to introduce a prescribed robustness margin $\eta_i > 0$, the watermark parameters can be selected such that

$$-p_{k,i}^\delta + g_{k,i} < \underline{r}_i - \eta_i$$

or

$$-p_{k,i}^\delta - g_{k,i} > \overline{r}_i + \eta_i.$$

A larger margin improves robustness against quantization errors, synchronization errors, and modeling inaccuracies. However, an unnecessarily large watermark amplitude increases the dynamic range of the transmitted residual and may require more quantization bits or communication bandwidth. Hence, in practical implementation, the watermark amplitude should be chosen as the smallest value that satisfies the above separation condition with the desired margin. Moreover, the watermark sequence itself does not need to be transmitted, since it is regenerated locally at the receiver using the shared PRNG seed. Thus, the additional communication overhead mainly comes from the increased dynamic range of the watermarked residual, rather than from transmitting extra watermark packets.

Remark 6. *The recursive construction of the error zonotope in the UIO may cause the number of generators to grow over time. To maintain computational efficiency in practical implementations, a reduction step (e.g., using the singular value decomposition (SVD)-based method in [48]) can be applied periodically to the watermark generator matrix H_k^Δ without losing the guaranteed inclusion property required by Theorem 2.*

Consider an adversary who intercepts the communication link and replaces the transmitted signal $\tilde{r}_k$ with a malicious value $\tilde{r}_k^a$. The recovered residual becomes:

$$\hat{r}_k^a = \tilde{r}_k^a - \delta_k. \tag{35}$$

Because δ_k is unknown to the adversary and varies with time, the attacker cannot compensate for it. Even a simple attack, such as setting $\tilde{r}_k^a = 0$ (i.e., a zero-value injection), results in $\hat{r}_k^a = -\delta_k$. If the watermark parameters are designed according to Theorem 2, every attack step will produce a recovered residual that violates the interval $[\underline{r}_i, \overline{r}_i]$, thereby guaranteeing detection. In general, any attempt to manipulate the data will be amplified by the unknown watermark, rendering the attack readily detectable.

3.3. MITM Attack Model

In this section, a MITM attacker model is introduced, wherein the adversary is assumed to have the capability to intercept and modify data transmitted over the wireless network between the smart sensor and the remote estimator. The attacker's goal is to degrade the state estimation performance while remaining undetected by the intrusion detector. It is assumed that the attacker can eavesdrop on the communication channel and replace $\tilde{r}_k$ with an arbitrary malicious value $\tilde{r}_k^a$. Moreover, the attacker does not have access to the internal states of the system, the observer, or the secret seed for watermark generation, but may have knowledge of the system matrices A, B, C, E, F and the noise bounds H_w, H_v (worst-case scenario).

Mathematically, the attack is modeled as

$$\tilde{r}_k^{\text{rcv}} = \begin{cases} \tilde{r}_k, & \text{attack-free at } k, \\ \tilde{r}_k^a, & \text{attacked at } k, \end{cases} \tag{36}$$

where $\tilde{r}_k^{\text{rcv}}$ denotes the signal received by the remote estimator. This paper focuses on a specific class of MITM attacks: zero-value injection attacks, in which the attacker sets $\tilde{r}_k^a = 0$ for all $k \geq k_{\text{attack}}$. This simple yet powerful attack can completely evade the residual-based detector in the absence of watermarking, as demonstrated in Section 4.

3.4. Detector

In this section, a residual-based detector is developed at the remote estimator to monitor the recovered residual $\hat{r}_k$ obtained from (27) after watermark removal. To distinguish between nominal system behavior and malicious cyber attacks, it is critical to construct a precise analytical bound for the residual under attack-free conditions, rather than relying on empirical simulations.

In the absence of attacks, the recovered residual perfectly matches the true residual, i.e., $\hat{r}_k = r_k = y_k - C\hat{x}_k$. According to the error dynamics defined in Section 3.1, the residual can be rewritten as $r_k = Ce_k + Fv_k$. Since the estimation error and the measurement noise are strictly bounded by zonotopes $e_k \in \mathcal{E}_k = \langle p_k^e, H_k^e \rangle$ and $v_k \in \mathcal{V} = \langle 0, H_v \rangle$, the nominal residual is guaranteed to be enclosed by a dynamically propagated zonotope $\mathcal{R}_k$:

$$r_k \in \mathcal{R}_k = \langle p_k^r, H_k^r \rangle,$$

where $p_k^r = Cp_k^e$ and $H_k^r = [CH_k^e, FH_v]$.

By computing the interval hull (i.e., Box($\mathcal{R}_k$)) of the residual zonotope, the exact mathematical lower and upper bounds for the i-th sensor channel at any time step k can be explicitly determined as:

$$\underline{r}_{k,i} = p_{k,i}^r - \sum_j |(H_k^r)_{ij}|,$$

$$\overline{r}_{k,i} = p_{k,i}^r + \sum_j |(H_k^r)_{ij}|.$$

To establish a robust, time-invariant detection threshold that completely eliminates false alarms without requiring empirical tolerance, the static threshold bounds $[\underline{r}_i, \overline{r}_i]$ are designed by taking the steady-state extremum of the dynamic bounds over all $k \geq 1$:

$$\underline{r}_i = \inf_{k \geq 1} \underline{r}_{k,i}, \quad \overline{r}_i = \sup_{k \geq 1} \overline{r}_{k,i}.$$

At each time step k, the detector compares the recovered residual $\hat{r}_k$ with these theoretical thresholds. The alarm is triggered according to the following logic:

$$\text{Alarm} = \begin{cases} 1, & \text{if } \hat{r}_{k,i} < \underline{r}_i \text{ or } \hat{r}_{k,i} > \overline{r}_i \text{ for any } i, \\ 0, & \text{otherwise.} \end{cases}$$

Because these thresholds are analytically derived using set-membership theory, they strictly encapsulate all the possible UBB noise realizations, mathematically ensuring a strictly zero false-alarm rate under nominal operating conditions. If the system is under a zero-value injection MITM attack, the watermarking scheme yields $\hat{r}_k = -\delta_k$. As proved in the watermark design (see the analysis in Section 3.2 and Theorem 2), the strategically configured watermark zonotope deterministically falls outside the nominal thresholds $[\underline{r}_i, \overline{r}_i]$, thereby triggering the alarm effectively.

To obtain a rigorous enclosure of e_k that is necessary for bounding the residual, the reachability analysis provided in Theorem 3 is employed. That theorem recursively propagates a zonotope

$$E_k = \langle p_k^e, H_k^e \rangle,$$

that strictly contains the estimation error for all $k \geq 0$, under the given unknown-but-bounded noise assumptions.

Based on the derived analytical thresholds, a proactive attack detection algorithm (Algorithm 1) utilizing dynamic zonotopic watermarking is proposed. By deliberately setting the watermark amplitude to strictly exceed the nominal thresholds, this mechanism ensures zero false alarms under attack-free conditions, while structurally forcing the recovered residual to violate the detection bounds under stealthy data manipulations (e.g., zero-value injections).

To rigorously guarantee that the estimated intervals enclose the true states at each time step, the following theorem establishes the analytical propagation of the error reachable set via zonotope operations.

Algorithm 1 Proactive attack detection based on dynamic watermarking

Require: Analytical thresholds $[\underline{r}^{th}, \overline{r}^{th}]$, shared PRNG seed
Ensure: Detection flag d_k
Offline Phase (Watermark Parameter Design):
Choose p_k^δ and H_k^δ such that the separation condition (28) or (29) is satisfied with a prescribed margin.
Online Phase (At each time step k):
Sensor Side
Generate $\delta_k = p_k^\delta + H_k^\delta z_k$ using the shared seed.
Transmit $\tilde{r}_k = (y_k - C\hat{x}_k) + \delta_k$
Remote Estimator Side
Receive $\tilde{r}_k^{rcv}$ and regenerate δ_k
Recover residual: $\hat{r}_k = \tilde{r}_k^{rcv} - \delta_k$
if $\exists i, \hat{r}_{k,i} < \underline{r}_i^{th}$ OR $\hat{r}_{k,i} > \overline{r}_i^{th}$ **then**
    $d_k = 1$ {Attack detected}
else
    $d_k = 0$ {Attack-free}
end if

Theorem 3. *For the system (1), the interval estimation of the estimation error $[\underline{e}_k, \overline{e}_k]$ can be obtained with $\hat{x}_0 = p_0$. When $k = 0$, the interval estimation of the estimation error e_0 can be obtained as*

$$[\underline{e}_0, \overline{e}_0] = Box(\langle 0, H_0\rangle). \tag{37}$$

When $k = 1$, the interval estimation of the estimation error e_1 can be represented as

$$[\underline{e}_1, \overline{e}_1] = Box(\Phi\langle 0, H_0\rangle) \oplus Box(-LF\mathcal{V}). \tag{38}$$

When $k \geq 2$, the interval estimation of e_k can be derived as

$$[\underline{e}_k, \overline{e}_k] = Box(\Phi^k\langle 0, H_0\rangle) \oplus \bigoplus_{i=0}^{k-1} Box(-\Phi^{k-1-i}LF\mathcal{V}). \tag{39}$$

Proof. Based on $x_0 \in \langle p_0, H_0\rangle$ and $\hat{x}_0 = p_0$, the initial estimation error is given by $e_0 = x_0 - \hat{x}_0 \in \langle 0, H_0\rangle$.

When $k = 0$, the estimation error is $e_0 = x_0 - \hat{x}_0$. The reachability set of the estimation error e_k is defined as $\mathcal{E}_k$. Then, when $k = 0$, $\mathcal{E}_0$ can be obtained as

$$\mathcal{E}_0 = \langle 0, H_0\rangle. \tag{40}$$

According to the error system (12) and the decoupling condition $NF = 0$, the error dynamics can be expressed as $e_k = \Phi e_{k-1} - LFv_{k-1}$. When $k = 1$, it gives

$$e_1 = \Phi e_0 - LFv_0. \tag{41}$$

Then, the reachable set $\mathcal{E}_1$ can be described as

$$\mathcal{E}_1 = \Phi\langle 0, H_0\rangle \oplus (-LF)\mathcal{V}. \tag{42}$$

When $k = 2$, by substituting the expression of e_1 into the error system, the error e_2 is expanded as

$$\begin{aligned} e_2 &= \Phi e_1 - LFv_1 \\ &= \Phi(\Phi e_0 - LFv_0) - LFv_1 \\ &= \Phi^2 e_0 - \Phi LFv_0 - LFv_1. \end{aligned} \tag{43}$$

By recursively expanding the above expression for $k \geq 2$, the general form of the estimation error e_k can be shown as

$$e_k = \Phi^k e_0 - \sum_{i=0}^{k-1} \Phi^{k-1-i} LFv_i. \tag{44}$$

Since $e_0 \in \mathcal{E}_0$ and $v_i \in \mathcal{V}$ [38], the reachable set $\mathcal{E}_k$ can be achieved using the properties of the Minkowski sum as

$$\mathcal{E}_k = \Phi^k \langle 0, H_0 \rangle \oplus \bigoplus_{i=0}^{k-1} (-\Phi^{k-1-i} LF)\mathcal{V}. \tag{45}$$

Subsequently, the interval estimation of the estimation error is obtained such that $[\underline{e}_k, \overline{e}_k] = \text{Box}(\mathcal{E}_k)$. Based on the relationship $x_k = \hat{x}_k - e_k$, the interval estimation of the system state is finally bounded by

$$\underline{x}_k = \hat{x}_k + \underline{e}_k, \quad \overline{x}_k = \hat{x}_k + \overline{e}_k. \tag{46}$$

The proof is completed. □

Remark 7. *The recursive construction of H_k^e may cause the number of columns to grow linearly with k. To maintain computational efficiency, a reduction step (e.g., using the method in [48]) can be applied periodically without losing the guaranteed inclusion property.*

Remark 8. *The exact decoupling of the process disturbance w_{k-1} is contingent upon the rank matching condition specified in Section 3.1. In practical applications where this condition is not strictly satisfied, the proposed UIO framework structurally degrades to a conventional Luenberger-type interval observer by setting $T = I_{n_x}$ and $N = 0$. Consequently, the influence of w_{k-1} is not eliminated but is instead robustly encapsulated within the interval estimation by augmenting the generator matrix update as $H_k^e = [\Phi H_{k-1}^e, EH_w, -LFH_v]$. While this suboptimal alternative preserves the guaranteed inclusion property, the resulting interval expansion significantly widens the detection thresholds, thereby increasing the system's susceptibility to stealthy attacks. This theoretical transition serves as the foundation for the comparative baseline analysis presented in Section 4.*

The overall procedure for interval estimation is summarized in Algorithm 2. Based on the designed UIO and zonotope reachability analysis, this algorithm recursively propagates the reachable set of the estimation error. By computing the interval hull of these dynamic zonotopes, it provides guaranteed upper and lower bounds for the system states at each time step.

Algorithm 2 Interval estimation

Require: $u_k, y_k, A, B, C, E, F, H_0, H_w, H_v$
Ensure: $\overline{x}_k, \underline{x}_k$
Initialization:
$X_0 = \langle 0, H_0 \rangle$, $\hat{x}_0 = p_0$
$N_0 = \emptyset$, $S^0 = -LF\mathcal{V}$
$[\underline{e}_0, \overline{e}_0] = \text{Box}(X_0)$
$\underline{x}_0 = \hat{x}_0 + \underline{e}_0, \quad \overline{x}_0 = \hat{x}_0 + \overline{e}_0$
Iterating for $k \geq 1$:
$X_k = \Phi X_{k-1}$ {Initial state attenuation}
$N_k = N_{k-1} \oplus S^{k-1}$ {Noise accumulation}
$S^k = \Phi S^{k-1}$
$[\underline{e}_0, \overline{e}_0] = \text{Box}(X_k \oplus N_k)$ {Error hull calculation}
$\hat{x}_k = TA\hat{x}_{k-1} + TBu_{k-1} + L(y_{k-1} - C\hat{x}_{k-1}) + Ny_k$
$\overline{x}_k = \hat{x}_k + \overline{e}_k, \quad \underline{x}_k = \hat{x}_k + \underline{e}_k$

Remark 9. *The computational burden of the proposed method can be divided into offline and online parts. In the offline stage, the rank condition verification, Moore-Penrose pseudoinverse calculation, LMI optimization in Theorem 1, watermark parameter design, and steady-state analytical threshold construction are performed only once before implementation. Therefore, these computations do not affect the real-time execution. In the online stage, the main operations at each sampling instant include the UIO update, residual computation, PRNG-based watermark generation/removal, and threshold comparison. Since the observer matrices, watermark parameters, and detection thresholds are precomputed, the online computation mainly involves fixed-size matrix-vector multiplications, with* *complexity approximately $O(n_x^2 + n_x n_u + n_x n_y)$. The watermark generation requires $O(n_y n_g)$ operations, which reduces to $O(n_y)$ when a diagonal generator matrix is used, and the threshold comparison requires only $O(n_y)$ operations.*

Moreover, the zonotopic operations used in the interval estimation are computationally attractive because zonotopes are closed under linear transformations and Minkowski sums. Although the number of generators may increase during recursive propagation, standard zonotope order-reduction techniques can be applied to keep the generator order bounded. Therefore, the online computational burden remains manageable, making the proposed method suitable for real-time implementation in CPSs with moderate state and output dimensions.

4. Simulation

In this section, the effectiveness of the proposed proactive defense framework is evaluated through numerical simulations on a second-order linear system. All system matrices, noise bounds, observer gains, and watermark parameters are explicitly provided to ensure reproducibility.

Consider the discrete-time linear system described by (1) with UBB process disturbances and measurement noises. The system matrices are chosen as

$$A = \begin{bmatrix} 0.8 & 0.1 \\ -0.2 & 0.9 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$E = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, F = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}.$$

The control input is set to $u_k = 0.8\sin(0.15k)$. The process noise $w_k \in \mathbb{R}$ and measurement noise $v_k \in \mathbb{R}$ are unknown but bounded, with known zonotopic bounds:

$$w_k \in \mathcal{W} = \langle 0, H_w \rangle, \qquad v_k \in \mathcal{V} = \langle 0, H_v \rangle,$$

where $H_w = 0.25$ and $H_v = 0.45$. The initial state is bounded within $\mathcal{X}_0 = \langle p_0, H_0 \rangle$ with $p_0 = [0,0]^T$, $H_0 = \mathrm{diag}([0.8, 0.8])$, and $x_0 = [0.4, -0.4]^T$.

For the proposed method, the rank condition $\mathrm{rank}([CE\ F]) = \mathrm{rank}([[CE\ F];[E\ 0]])$ was satisfied, yielding exact disturbance decoupling via $N = [1, -0.5; 0, 0]$ and $T = [0, 0.5; 0, 1]$, while $NF = 0$. The observer gain L was obtained via Theorem 1 by solving the LMI (15), yielding the optimal H_∞ attenuation level $\gamma = 2.52 \times 10^{-3}$ and

$$L = \begin{bmatrix} -0.2600 & 0.1300 \\ -0.5200 & 0.2600 \end{bmatrix},$$

which results in the stable error dynamics matrix $\Phi = TA - LC$ with eigenvalues $\{0, 0.8\}$. The baseline Luenberger observer is designed by pole placement at $\{0.5, 0.6\}$ with $T = I_2$, $N = 0$, serving as the structural degradation when both decoupling conditions fail.

Analytical thresholds were derived via 100-step steady-state zonotope propagation. For the baseline, the error generator accumulated uncompensated uncertainty EH_w at each recursion, producing conservative bounds $[\underline{r}_1, \overline{r}_1] = [-0.5573, 0.5573]$. The UIO restricted the residual set to FH_v and converging error dynamics, yielding tight bounds $[-0.2250, 0.2250]$ for channel 1.

A zero-value injection attack is launched from $k = 30$ to $k = 50$. The watermark was designed according to Theorem 2 with $H_k^{\Delta} = 0.1 I_2$ and $p_k^{\Delta} = [0.3750, 0.6000]^T$, satisfying (29) with a safety margin of 0.05. Under attack, the recovered residual evaluates to $\hat{r}_k = -\delta_k \in [-0.4750, -0.2750] \times [-0.7000, -0.5000]$, which strictly falls outside the nominal thresholds.

Figures 1 and 2 compare the state estimation performance. The baseline interval is excessively wide due to uncompensated disturbance accumulation [23], whereas the UIO achieves tight bounds via the exact decoupling conditions $TE = 0$ and $NF = 0$. The identical estimates prior to $k = 30$ verify watermark transparency according to Remark 3, confirming that the dynamic watermark does not degrade nominal estimation accuracy.

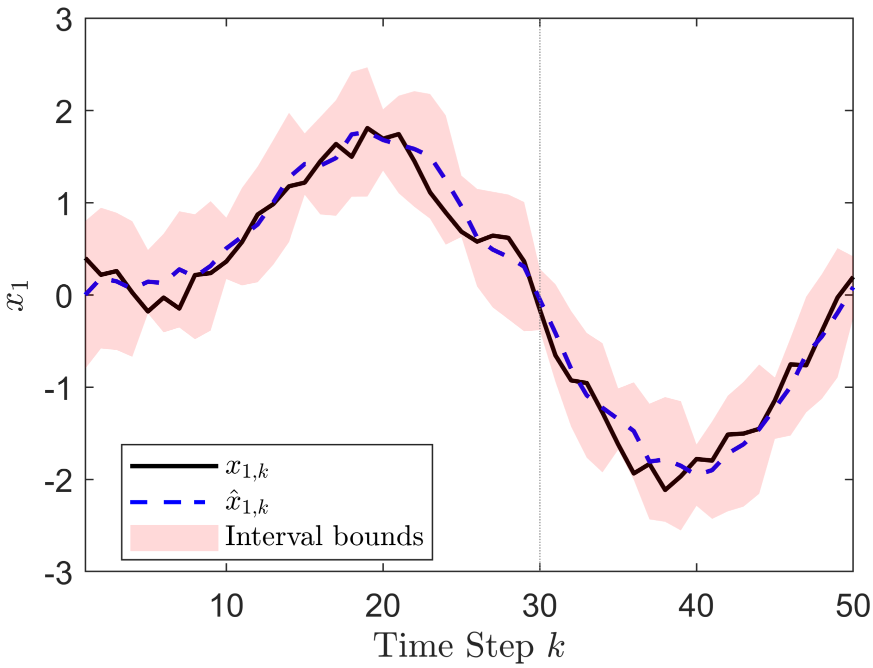

Figure 1. Baseline: state $x_{1,k}$ (black), estimate $\hat{x}_{1,k}$ (dashed), interval bounds (shaded).

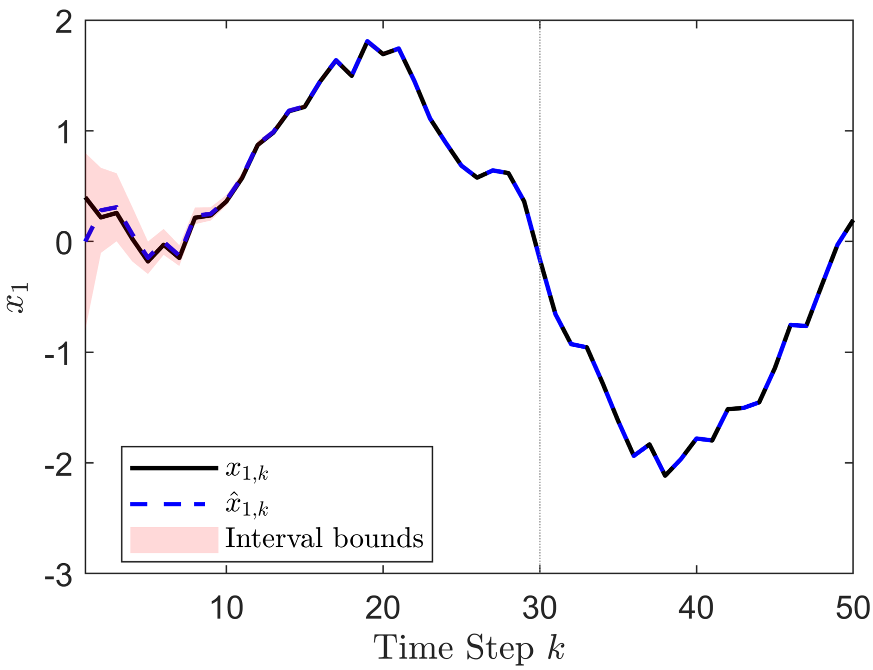

Figure 2. Proposed: state $x_{1,k}$ (black), estimate $\hat{x}_{1,k}$ (dashed), tight interval bounds (shaded).

Figures 3 and 4 compare the residual behavior under attack. The baseline residual $\hat{r}_{1,k} = 0$ remains within $[-0.5573, 0.5573]$ (Figure 3), while the proposed recovered residual $\hat{r}_{1,k} = -\delta_{1,k}$ violates $[-0.2250, 0.2250]$ at every step (Figure 4). This contrast arises because the baseline threshold is too conservative to expose zero-value injections, whereas the watermark is deliberately designed according to Theorem 2 to force the recovered residual outside the nominal bounds. Consequently, the proactive defense achieves guaranteed detection with zero false alarms.

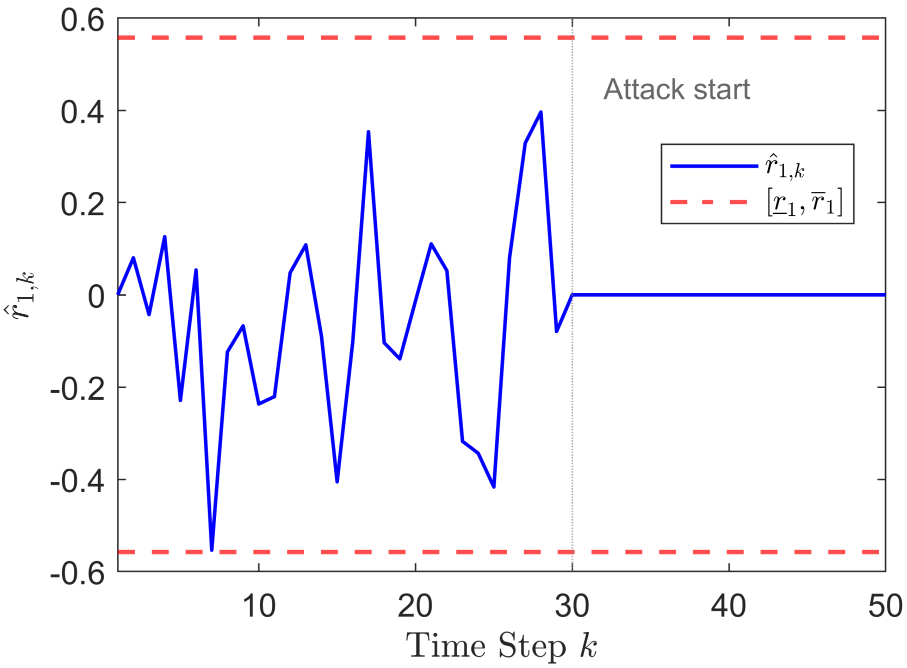

Figure 3. Baseline: received residual $\hat{r}_{1,k}$ (blue) within thresholds $[-0.5573, 0.5573]$ (dashed red).

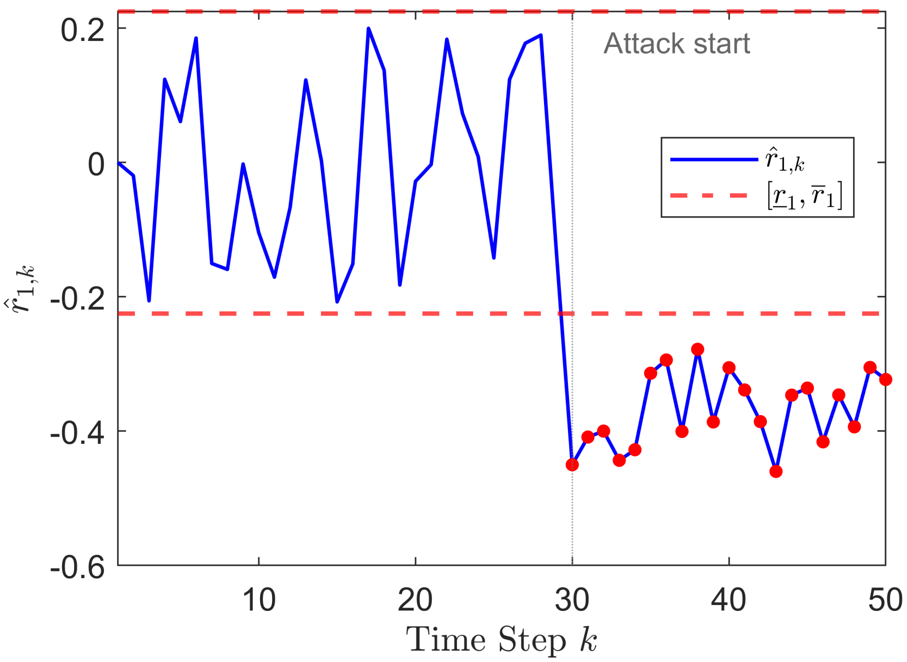

Figure 4. Proposed: recovered residual $\hat{r}_{1,k}$ (blue) violating thresholds $[-0.2250, 0.2250]$ (dashed red).

Figures 5 and 6 compare the detection flags. The baseline yields zero alarms throughout the attack horizon (Figure 5), while the proposed method triggers at every attack step (Figure 6). This discrepancy stems from the fact that the baseline threshold is too wide to expose zero-value injections, whereas the proposed watermarking scheme strictly satisfies the deterministic detection condition in Theorem 2. As a result, the proactive defense achieves a strictly zero false-alarm rate under nominal conditions while guaranteeing 100% detection against stealthy MITM attacks.

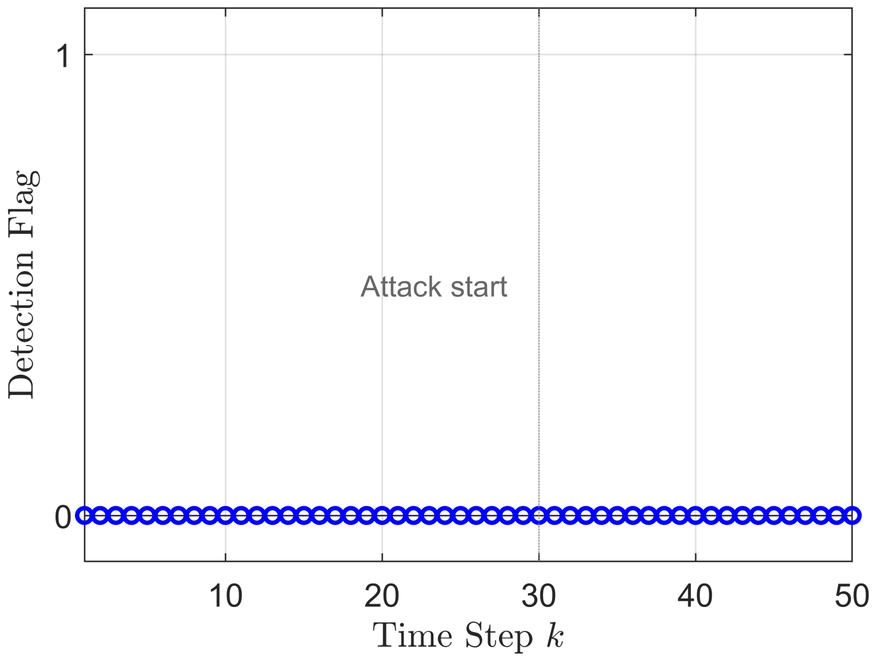

Figure 5. Baseline: detection flag (0% detection rate).

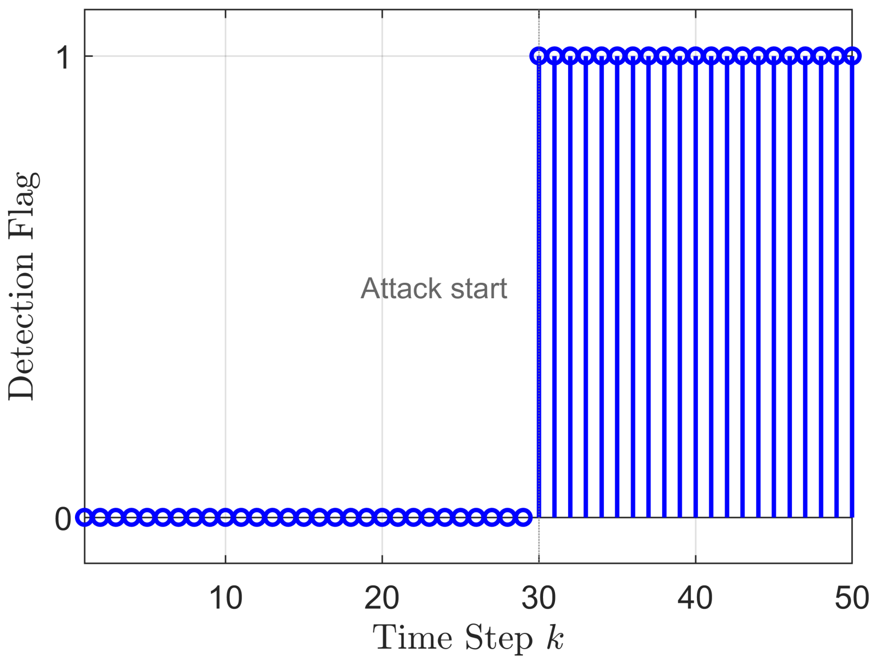

Figure 6. Proposed: detection flag (100% detection rate).

Table 1 reports the quantitative detection statistics. The attack remains undetected by the baseline, as the injected zero value resides within its conservative threshold band. The proposed scheme achieves deterministic detection with strictly zero false alarms under nominal conditions, validating Theorem 2 and Remark 3.

Table 1. Detection performance comparison.

Method	False Alarms ($k < 30$)	Detected ($k \geq 30$)	Rate
Baseline	0	0/21	0%
Proposed	0	21/21	100%

5. Conclusions

In this work, an improved proactive defense strategy has been proposed for secure remote state estimation in CPSs subject to UBB noises, disturbances, and MITM attacks. By integrating a UIO with a decoupling approach and H_∞ optimization, the proposed method achieves high-accuracy interval estimation with the true states strictly bounded within dynamically propagated zonotopes. Furthermore, a time-varying unpredictable zonotopic watermarking mechanism ensures strict transparency under attack-free conditions while deterministically exposing stealthy data manipulations. Numerical simulations demonstrate that the proposed strategy guarantees reliable detection of zero-value injection attacks with a strictly zero false-alarm rate, validating the effectiveness of the proactive defense framework. Future work will extend the proposed framework to nonlinear CPSs by incorporating nonlinear observer design and zonotopic reachability analysis. Experimental validation on practical networked control platforms will also be investigated.

Author Contributions

Z.L.: Conceptualization, methodology, software, validation, formal analysis, investigation, writing original draft, and visualization; C.L.: Methodology, validation, data analysis, and writing review and editing; Z.W.: Supervision, project administration, funding acquisition, conceptual guidance, and writing review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Shandong Province Natural Science Foundation under the ZR2025MS1037 Grant.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Ding, D.; Han, Q.-L.; Ge, X.; et al. Secure State Estimation and Control of Cyber-Physical Systems: A Survey. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *51*, 176–190.
2. Guo, Z.; Shi, D.; Johansson, K.H.; et al. Worst-Case Stealthy Innovation-Based Linear Attack on Remote State Estimation. *Automatica* **2018**, *89*, 117–124.
3. Huang, J.; Ho, D.W.C.; Li, F.; et al. Secure Remote State Estimation Against Linear Man-in-the-Middle Attacks Using Watermarking. *Automatica* **2020**, *121*, 109182.
4. Zhang, Y.; Peng, Z.; Wen, G.; et al. Optimal Stealthy Linear Man-in-the-Middle Attacks with Resource Constraints on Remote State Estimation. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 445–456.
5. Jovanov, I.; Pajic, M. Relaxing Integrity Requirements for Attack-Resilient Cyber-Physical Systems. *IEEE Trans. Autom. Control* **2019**, *64*, 4843–4858.
6. Gao, Y.; Sun, G.; Liu, J.; et al. State Estimation and Self-Triggered Control of CPSs against Joint Sensor and Actuator Attacks. *Automatica* **2020**, *113*, 108687.
7. Guo, Z.; Wang, X.; Tian, M.; et al. Zonotopic State Estimation and Fault Detection for Discrete-Time Linear Systems Under DoS Attacks: A Switching Controller Design Mechanism. *Automatica* **2026**, *185*, 112833.
8. Hu, C.; Ding, S.; Xie, X. Event-Based Distributed Set-Membership Estimation for Complex Networks under Deception Attacks. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 3719–3729.
9. Shoukry, Y.; Nuzzo, P.; Puggelli, A.; et al. Secure State Estimation for Cyber-Physical Systems Under Sensor Attacks: A Satisfiability Modulo Theory Approach. *IEEE Trans. Autom. Control* **2017**, *62*, 4917–4932.
10. Hu, S.; Chen, X.; Li, J.; et al. Observer-Based Resilient Controller Design for Networked Stochastic Systems under Coordinated DoS and FDI Attacks. *IEEE Trans. Control Netw. Syst.* **2024**, *11*, 890–901.
11. Zhu, K.; Wang, Z.; Ding, D.; et al. Encryption-Decryption-Based Set-Membership Filtering for Two-Dimensional Systems: On Security and Boundedness. *Automatica* **2025**, *173*, 112091.

12. Hu, C.; Xie, X.; Ding, S.; et al. Distributed Set-Membership Fusion Estimation for Complex Networks with Communication Constraints. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 3877–3886.
13. Casini, M.; Garulli, A.; Vicino, A. Set Membership State Estimation for Discrete-Time Linear Systems with Binary Sensor Measurements. *Automatica* **2024**, *159*, 111396.
14. Yang, H.; Yan, H.; Li, Z.; et al. Set-Membership Estimation for Nonlinear Systems Based on Zonotope Analysis. *Automatica* **2025**, *177*, 112278.
15. Wang, Z.; Lian, D.; Puig, V.; et al. Fault Detection for Brushless Direct-Current Motor Using Descriptor System-Based Set-Membership Estimation. *IEEE Trans. Control Syst. Technol.* **2025**, *33*, 1640–1650.
16. Efimov, D.; Raissi, T.; Zolghadri, A. Control of Nonlinear and LPV Systems: Interval Observer-Based Framework. *IEEE Trans. Autom. Control* **2013**, *58*, 773–778.
17. Li, G.; Wang, Z.; Bai, X.; et al. Dynamic Event-Triggered State Estimation for Power Harmonics with Quantization Effects: A Zonotopic Set-Membership Approach. *IEEE Trans. Sustain. Comput.* **2024**, *9*, 803–813.
18. Espindola-Winck, G.; Hardouin, L.; Lhommeau, M. On the Set-Estimation of Uncertain Max-Plus Linear Systems. *Automatica* **2025**, *171*, 111899.
19. Han, W.; Wang, Z.; Shen, Y.; et al. Interval Estimation for Uncertain Systems via Polynomial Chaos Expansions. *IEEE Trans. Autom. Control* **2021**, *66*, 468–475.
20. Yang, H.; Yan, H.; Shi, Y.; et al. Distributed Zonotopic State Estimation for 2-D Systems over Wireless Sensor Networks. *Automatica* **2026**, *190*, 113070.
21. Fei, Z.; Yang, L.; Sun, X.-M.; et al. Zonotopic Set-Membership State Estimation for Switched Systems with Restricted Switching. *IEEE Trans. Autom. Control* **2022**, *67*, 6127–6134.
22. Kulmburg, A.; Schäfer, L.; Althoff, M. Approximability of the Containment Problem for Zonotopes and Ellipsotopes. *IEEE Trans. Autom. Control* **2025**, *70*, 8104–8119.
23. Liang, C.; Niu, B.; Guo, Z.; et al. Interval Estimation for Discrete-Time Linear Systems Based on Unknown Input Observer and Reachability Analysis. *Asian J. Control* **2026**, *28*, 773–781.
24. Yang, G.; Rezaee, H.; Serrani, A.; et al. Sensor Fault-Tolerant State Estimation by Networks of Distributed Observers. *IEEE Trans. Autom. Control* **2022**, *67*, 5348–5360.
25. Xing, M.; Lu, J.; Liu, Y. Remote Secure Fusion Estimation of Cyber–Physical Systems under False Data Injection Attacks. *Automatica* **2025**, *177*, 112271.
26. Liu, Q.; Wang, J.; Ni, Y.; et al. Performance Analysis for Cyber–Physical Systems under Two Types of Stealthy Deception Attacks. *Automatica* **2024**, *160*, 111446.
27. Liu, H.; Wang, S.; Niu, B.; et al. A Novel Attack Detection for Linear Systems with Unknown-But-Bounded Noises. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 4223–4232.
28. Keshk, M.; Sitnikova, E.; Moustafa, N.; et al. An Integrated Framework for Privacy-Preserving Based Anomaly Detection for Cyber-Physical Systems. *IEEE Trans. Sustain. Comput.* **2021**, *6*, 66–79.
29. Samineni, N.R.; Barbhuiya, F.A.; Nandi, S. Stealth and Semi-Stealth MITM Attacks, Detection and Defense in IPv4 Networks. In Proceedings of the 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, Solan, India, 6–8 December 2012; pp. 364–367.
30. Liang, C.; Guo, Z.; Niu, B.; et al. Sensor Attacks Detection and Reconstruction for AUVs: An Improved Zonotopic Analysis Approach. *IEEE Internet Things J.* **2025**, *12*, 40620–40627.
31. Liang, C.; Niu, B.; Guo, Z.; et al. Attack Detection and Reconstruction for CPSs Based on Unknown Input Observer and Reachability Analysis. *IEEE Trans. Syst. Man Cybern. Syst.* **2025**, *55*, 9645–9654.
32. Zhang, K.; Keliris, C.; Parisini, T.; et al. Passive Attack Detection for a Class of Stealthy Intermittent Integrity Attacks. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 898–915.
33. Pasqualetti, F.; Dörfler, F.; Bullo, F. Attack Detection and Identification in Cyber-Physical Systems. *IEEE Trans. Autom. Control* **2013**, *58*, 2715–2729.
34. Liu, H.; Li, Y.; Han, Q.-L.; et al. Watermark-Based Proactive Defense Strategy Design for Cyber-Physical Systems with Unknown-but-Bounded Noises. *IEEE Trans. Autom. Control* **2023**, *68*, 3300–3315.
35. Porter, M.; Hespanhol, P.; Aswani, A.; et al. Detecting Generalized Replay Attacks via Time-Varying Dynamic Watermarking. *IEEE Trans. Autom. Control* **2021**, *66*, 3502–3517.
36. Hespanhol, P.; Porter, M.; Vasudevan, R.; et al. Sensor Switching Control Under Attacks Detectable by Finite Sample Dynamic Watermarking Tests. *IEEE Trans. Autom. Control* **2021**, *66*, 4560–4574.
37. Shangguan, L.; Chour, K.; Ko, W.H.; et al. Dynamic Watermarking for Cybersecurity of Autonomous Vehicles. *IEEE Trans. Ind. Electron.* **2023**, *70*, 11735–11743.
38. Combastel, C. Zonotopes and Kalman Observers: Gain Optimality under Distinct Uncertainty Paradigms and Robust Convergence. *Automatica* **2015**, *55*, 265–273.
39. Tang, W.; Wang, Z.; Shen, Y. Interval Estimation for Discrete-Time Linear Systems: A Two-Step Method. *Syst. Control Lett.* **2019**, *123*, 69–74.
40. Qing, H. Set Membership State Estimation for Nonlinear Systems in the Presence of Bounded Disturbances. In Proceedings

of the 2007 Chinese Control Conference, Zhangjiajie, China, 26–31 July 2007; pp. 196–201.
41. Silvestre, D. Comparison of Recent Advances in Set-Membership Techniques: Application to State Estimation, Fault Detection and Collision Avoidance. *Eur. J. Control* **2025**, *81*, 101161.
42. Xi, R.; Tang, W.; Chen, J. Robust Sliding Mode Control Based on Set-Membership Estimation for UAV Trajectory Tracking Under External Disturbances. In Proceedings of the 14th IEEE Data Driven Control and Learning Systems (DDCLS), Wuxi, China, 9–11 May 2025; pp. 1008–1013.
43. Zhang, F., ed. *The Schur Complement and Its Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
44. Barata, J.C.A.; Hussein, M.S. The Moore-Penrose Pseudoinverse: A Tutorial Review of the Theory. *Braz. J. Phys.* **2012**, *42*, 146–165.
45. Guan, Y.; Saif, M. A Novel Approach to the Design of Unknown Input Observers. *IEEE Trans. Autom. Control* **1991**, *36*, 632–635.
46. Cao, G.; Wang, J. Distributed Unknown Input Observer. *IEEE Trans. Autom. Control* **2023**, *68*, 8244–8251.
47. Penrose, R. A Generalized Inverse for Matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*; Cambridge University Press: London, UK, 1955; Vol. 51, pp. 406–413.
48. Kopetzki, A.-K.; Schürmann, B.; Althoff, M. Methods for Order Reduction of Zonotopes. In Proceedings of the 56th IEEE Annual Conference on Decision and Control (CDC), Melbourne, VIC, Australia, 12–15 December 2017; pp. 5626–5633.