Model-based Randomness Monitor for Stealthy Sensor Attacks

Paul J Bonczek, Shijie Gao, and Nicola Bezzo

***Abstract*— Malicious attacks on modern autonomous cyber-physical systems (CPSs) can leverage information about the system dynamics and noise characteristics to hide while hijacking the system toward undesired states. Given attacks attempting to hide within the system noise profile to remain undetected, an attacker with the intent to hijack a system will alter sensor measurements, contradicting with what is expected by the system's model. To deal with this problem, in this paper we present a framework to detect non-randomness in sensor measurements on CPSs under the effect of sensor attacks. Specifically, we propose a run-time monitor that leverages two statistical tests, the *Wilcoxon Signed-Rank* test and *Serial Independence Runs* test to detect inconsistent patterns in the measurement data. For the proposed statistical tests we provide formal guarantees and bounds for attack detection. We validate our approach through simulations and experiments on an unmanned ground vehicle (UGV) under stealthy attacks and compare our framework with other anomaly detectors.**

I. Introduction

Modern autonomous systems are fitted with multiple sensors, computers, and networking devices that make them capable of many applications with little/no human supervision. Autonomous navigation, transportation, surveillance, and task oriented jobs are becoming more common and ready for deployment in real world applications especially in the automotive, industrial, and military domains. These enhancements in autonomy are possible thanks to the tight interaction between computation, sensing, communications, and actuation that characterize cyber-physical systems (CPSs). These systems are however vulnerable and susceptible to cyber-attacks like sensor spoofing which can compromise their integrity and the safety of the surroundings. In the context of autonomous vehicle technologies, one of the most typical threats is *hijacking* in which an adversary is capable to administer malicious attacks with the intent of leading the system to an undesired state. An example of this problem was demonstrated by authors in [1] in which GPS data were spoofed to slowly drive a yacht off the intended route.

If we look at the specific architecture of these robotic systems, typical autonomous applications employ go-to-goal and trajectory tracking and if one or more on-board sensors are compromised, system behavior can become unreliable. These vehicles typically have well studied dynamics and their sensors have specific expected behaviors according to their characterized noise models. An attacker that wants to perform a malicious hijacking can create non-random patterns or add small biases in the measurements to slowly push the system towards undesired states, for example creating undesired deviations as depicted in Fig. 1, all while remaining hidden within the system's and sensors noise profile. Hence, in order for an attacker to hijack the system with stealthy attack signals, a violation to the expected random behavior of the sensor measurements must occur.

Paul J Bonczek, Shijie Gao, and Nicola Bezzo are with the Charles L. Brown Department of Electrical and Computer Engineering, and Link Lab, University of Virginia, Charlottesville, VA 22904, USA. Email: {pjb4xn, sg9dn, nb6be}@virginia.edu

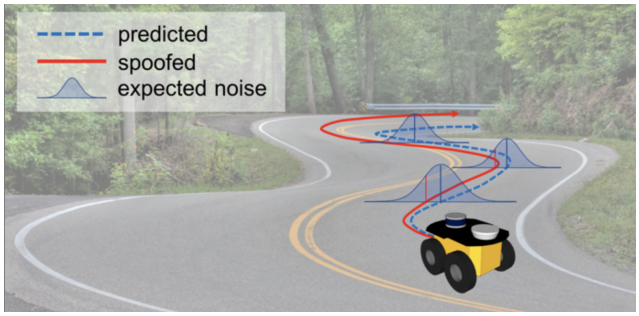

Fig. 1. A pictorial representation of the problem discussed in this paper in which a cyber-attack is able to hijack a vehicle into unsafe states while remaining hidden within the noise profile of its sensors.

With these considerations and problem in mind, in this work, we leverage the known characteristics of the *residual* – the difference between sensor measurements and state prediction – to build a run-time monitor to detect non-random behaviors. To monitor randomness, the non-parametric statistical Wilcoxon Signed-Rank [2] and Serial Independence Runs [3] tests are applied to individual sensors to determine if their measurements are being received randomly. The Wilcoxon test is an indicator of whether the residual is symmetric over its expected value, whereas the Serial Independence runs test indicates whether the sequences of residuals are arriving in a random manner. Thus, the main objective of this work is to find hidden attacks exhibiting non-random behavior within the noise. Given the nature of the non-parametric statistical tests that we propose, only random behavior of the residual is considered here, leaving the magnitude bounds of the residual un-monitored. Several detectors providing magnitude bounds on attacks have been already researched in the literature, thus in this work we also present a framework to combine existing approaches for magnitude bound detection with the proposed randomness monitor. In doing so, our approach improves the state-of-the-art attack detection by adding an extra layer of checks.

A. Related Work

This work builds on previous research considering deceptive cyber-attacks to *hijack* a system by injecting false data to sensor measurements while trying to remain undetected [4]. Many of the previous works use the residual for detection, which gives clues whether sensor measurements are healthy (uncompromised). Previous works characterizing the effects of stealthy sensor attacks on the Kalman filter can be found in [5], [6]. Similarly, authors in [4], [7] discuss how stealthy, undetectable attacks can compromise closed-loop systems, causing state and system dynamic degradation

Several procedures and techniques that analyze the residual for attack detection exist, one of which is the Sequential Probability Ratio Testing (SPRT) [8] that tests the sequence

of incoming residuals one at a time by taking the log-likelihood function (LLF). The Cumulative Sum (CUSUM) procedure proposed in [9] and [10] leverages the known characteristics of the residual covariance and sequentially sums the residual error to find changes in mean of the distribution. Compound Scalar Testing (CST) in [7] is another technique which is computationally friendly by reducing the residual vector with the known residual covariance matrix into a scalar value with χ^2 distribution. An improvement of CST in [11] is made by including a coding matrix to sensor outputs that is unknown to attackers, then an iterative optimization algorithm is used to solve for a transform matrix to detect stealthy attacks. Similar to our work where monitors are placed on individual sensors, the authors in [12] propose a Trust-based framework for sensor sets by "side-channel" monitors to provide a weight for trustworthiness to determine whether sensors have been compromised. Other works have proposed attack resiliency by leveraging information from redundant sensing. In [13], authors solve to reconstruct the state estimate of stochastic systems using an l_0 optimization problem when less than half of the sensors are compromised. Different from these previous works, we build a framework to monitor sensor measurements to find previously undetectable attacks by searching for non-random behavior.

The remainder of this work is organized as follows. In Section II we begin with system, estimator models and problem formulation, followed by the description of our Random Monitor framework in Section III. In Section IV an analysis of worst-case stealthy attacks and characterization of the effects on system performance is presented. Finally, in Section V we demonstrate through simulations and experiments the performance of our framework augmented with boundary detectors before drawing conclusions in Section VI.

II. Preliminaries & Problem Formulation

In this work we consider autonomous systems whose dynamics can be described by a discrete-time linear time-invariant (LTI) system in the following form:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{\nu}_k \\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\eta}_k, \end{aligned} \tag{1}$$

with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ the state matrix, $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ the input matrix, and $\boldsymbol{C} \in \mathbb{R}^{s\times n}$ the output matrix with the state vector $\boldsymbol{x}_k \in \mathbb{R}^n$, system input $\boldsymbol{u}_k \in \mathbb{R}^m$, output vector $\boldsymbol{y}_k \in \mathbb{R}^s$ providing measurements from s sensors from the set $\mathcal{S} = \{1, 2, \ldots, s\}$, and sampling time-instants $k \in \mathbb{N}$. Process and measurement noises are i.i.d. multivariate zero-mean Gaussian uncertainties $\boldsymbol{\nu} = \mathcal{N}(0, \boldsymbol{Q}) \in \mathbb{R}^n$ and $\boldsymbol{\eta} = \mathcal{N}(0, \boldsymbol{R}) \in \mathbb{R}^s$ with covariance matrices $\boldsymbol{Q} \in \mathbb{R}^{n\times n}, \boldsymbol{Q} \geq 0$ and $\boldsymbol{R} \in \mathbb{R}^{s\times s}, \boldsymbol{R} \geq 0$ and are assumed static.

During operations, a standard Kalman Filter (KF) is implemented to provide a state estimate $\hat{\boldsymbol{x}}_k \in \mathbb{R}^n$ in the form

$$\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{L}(\boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k), \tag{2}$$

where the Kalman gain matrix $\boldsymbol{L} \in \mathbb{R}^{n\times s}$ is

$$\boldsymbol{L} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^T(\boldsymbol{R} + \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T)^{-1}, \tag{3}$$

therefore, we assume that the KF is at steady state, i.e., $\lim_{k\to\infty} \boldsymbol{P}_k = \boldsymbol{P}$. The estimation error of the KF is defined as $\boldsymbol{e}_k = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k$ while its *residual* $\boldsymbol{r}_k$ is given by

$$\boldsymbol{r}_k = \boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k = \boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k, \tag{4}$$

The covariance of the residual (4) is defined as

$$\boldsymbol{\Sigma} = \mathrm{E}[\boldsymbol{r}_{k+1}\boldsymbol{r}_{k+1}^T] = \boldsymbol{R} + \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T \in \mathbb{R}^{s\times s}. \tag{5}$$

In the absence of sensor attacks, the residual for the i^{th} sensor $r_{k,i}, i \in \mathcal{S}$ follows a Gaussian distribution $r_{k,i} \sim \mathcal{N}(0, \sigma_i^2)$ where σ_i^2 is the i^{th} diagonal element of the residual covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{s\times s}$ in (5) such that

$$\mathrm{E}[r_{k,i}] = 0, \ \mathrm{Var}[r_{k,i}] = \sigma_i^2. \tag{6}$$

We describe the system output considering sensor attacks as

$$\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k, \tag{7}$$

where $\boldsymbol{\xi}_k \in \mathbb{R}^s$ represents the sensor attack vector. Our proposed framework consists in adding a monitor on each sensor searching for non-random behavior of the sensor measurement residual, hence any sensor may be compromised.

Definition 1: A sensor measurement is random if:

- a sequence of residuals over a time window occurs in an unpredictable, pattern-free manner.
- residuals have proper distributions over $\mathrm{E}[\boldsymbol{r}_k]$.

Since we are considering sensor spoofing, an attack signal $\boldsymbol{\xi}_k$ containing malicious data can disrupt randomness, causing measurements to display non-random behavior. Formally, the problem that we are interested in solving is:

Problem 1: **Randomness of Measurements:** Given the residual $\boldsymbol{r}_k$ between a measurement $\boldsymbol{y}_k$ and the corresponding prediction $\boldsymbol{C}\hat{\boldsymbol{x}}_k$ as defined in (4), find a policy to determine at run-time whether a sensor measurement is random, i.e., if any condition in Definition 1 does not hold.

III. Randomness Monitoring Framework

The overall cyber-physical system architecture including our Randomness Monitor framework is summarized in Fig. 2. The Randomness Monitor, augmented to any boundary detector providing magnitude bounds, is placed in the system feedback to monitor the residual sequence. We

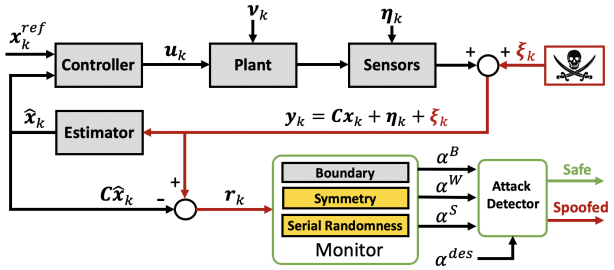

Fig. 2. The architecture of a CPS while experiencing sensor attacks augmented with our monitoring technique.

introduce a framework to monitor randomness of the residual sequence through two tests and provide tuning bounds for each to result in desired false alarm rates. From (4), the residual should have a symmetric distribution centered at zero and the sequence of residuals should arrive in a random order, having an absence of structure or patterns. For example, a continuously alternating pattern of "negative" and "positive" values, or a pattern of only "negative" values would clearly not satisfy random sequences.

Both tests operate online providing an alarm when the residual does not satisfy the conditions of each test. A desired false alarm rate $\alpha_i^{\text{des}} \in (0, 1)$ for each i^{th} sensor is tuned for

each test, and in the absence of sensor attacks, the observed alarm rate $\alpha_i \in [0,1]$ for each test should match closely with the tuned desired value $\alpha_i \sim \alpha_i^{\text{des}}$.

A. Residual Symmetry Monitor

To monitor whether the sequence of residuals are symmetrically distributed and zero-mean, we leverage the Wilcoxon Signed-Rank (WSR) test [2] as follows. A hypothesis test is formed by $\mathcal{H}_0$ for no attacks and $\mathcal{H}_a$ with attacks:

$$\begin{cases} \mathcal{H}_0 : \mathrm{E}[\boldsymbol{r}_k] = 0 \textbf{ and } \boldsymbol{r}_k \text{ is symmetric}, \\ \mathcal{H}_a : \mathrm{E}[\boldsymbol{r}_k] \neq 0 \quad \textbf{or } \boldsymbol{r}_k \text{ is not symmetric}. \end{cases} \tag{8}$$

A monitor is built to check if the residual $\boldsymbol{r}_k$ sequence over a sliding monitoring window $T = (k-\ell+1, k)$ for ℓ previous steps is symmetric. We denote the vector of residual sequences over the sliding window T as $\boldsymbol{r}_T = (\boldsymbol{r}_{T,1}, \ldots, \boldsymbol{r}_{T,i}, \ldots, \boldsymbol{r}_{T,s})$ where the residual sequence for an i^{th} sensor is $\boldsymbol{r}_{T,i} = (r_{k-\ell+1,i}, \ldots, r_{k,i})$. Following $\mathcal{H}_0$, we would expect that the number of positive and negative values of $\boldsymbol{r}_k$ over the monitoring window are equal. Additionally, a symmetric distribution indicates that the expected absolute magnitude of positive and negative residuals over a given window of length ℓ are equal,

$$\mathrm{E}[|\boldsymbol{r}_{T,i}^{+}|] = \mathrm{E}[|\boldsymbol{r}_{T,i}^{-}|], \ i \in \mathcal{S}, \tag{9}$$

where $\mathrm{E}[|\boldsymbol{r}_{T,i}^{+}|]$ and $\mathrm{E}[|\boldsymbol{r}_{T,i}^{-}|]$ denote the expected absolute magnitude for positive and negative values of the residual $r_{k,i}$ within the window T for any given i^{th} sensor. In other words, we would expect the sum of absolute values from the residual to be equal for both the positive and negative values. The WSR test takes both the sign and magnitude of the residual into account to determine whether conditions satisfy $\mathcal{H}_0$. Large differences in the residual signs or signed magnitudes imply non-similar distributions, causing the test to reject the no attack assumption and triggering an alarm.

To perform the WSR test at each time step k, we first look at the ℓ number of residuals over the monitoring window T of a given i^{th} sensor, ranking the *absolute values* of residuals $r_{T,i}$, starting with $rank = 1$ for the smallest absolute value, $rank = 2$ for the second smallest, and so on until reaching the largest absolute value with $rank = \ell$. Ranks of absolute values for positive (i.e. $|r_{T,i}^{+}|$) and negative (i.e. $|r_{T,i}^{-}|$) residuals over the window T are placed into the sets $\mathcal{R}_{k,i}^{+}$ and $\mathcal{R}_{k,i}^{-}$ at every time instance k, respectively.

Remark 1: For residuals equal to each other and not equal to 0 (tied for the same rank), an average of the ranks that would have been assigned to these residuals is given to each of the tied values. Furthermore, residuals equal to 0 are removed and ℓ is reduced accordingly.

Following, we compute the sum of ranks for both the positive and negative valued residuals,

$$W_{k,i}^{+} = \sum \mathcal{R}_{k,i}^{+}, \quad W_{k,i}^{-} = \sum \mathcal{R}_{k,i}^{-}. \tag{10}$$

Residuals with symmetric distributions have similar valued sum of ranks, i.e. $W_{k,i}^{+} \sim W_{k,i}^{-}$, whereas the sum of ranks in non-symmetric distributions are not similar $W_{k,i}^{+} \nsim W_{k,i}^{-}$ resulting in a rejection of $\mathcal{H}_0$ in (8), which we will now discuss how to solve. Assuming a large window of size $\ell \geq 20$[1] [14], the Wilcoxon random variables $W_{k,i}^{+}$, $W_{k,i}^{-}$ converge to a Normal distribution (without attacks) as $\ell \to \infty$ and can be approximated to a standard normal distribution. The approximated expected value and variance of the two sum of ranks $W_{k,i}^{+}$ and $W_{k,i}^{-}$, denoted as $W_{k,i}^{\pm} = \{W_{k,i}^{+}, W_{k,i}^{-}\}$ is

$$\mathrm{E}[W_{k,i}^{\pm}] = \tfrac{\ell^2+\ell}{4}, \quad \mathrm{Var}[W_{k,i}^{\pm}] = \tfrac{(\ell^2+\ell)(2\ell+1)}{24}. \tag{11}$$

The z-score of (10) for a given i^{th} sensor is computed by

$$Z_{k,i}^{W} = \frac{\min(W_{k,i}^{\pm}) - \mathrm{E}[W_{k,i}^{\pm}]}{\sqrt{\mathrm{Var}[W_{k,i}^{\pm}]}} = \frac{\min(W_{k,i}^{\pm}) - \frac{(\ell^2+\ell)}{4}}{\sqrt{\frac{(\ell^2+\ell)(2\ell+1)}{24}}}, \tag{12}$$

and the p-value used to determine whether to reject the null hypothesis $\mathcal{H}_0$ (i.e. no attacks) is computed from (12) as

$$p_{k,i}^{W} = \Phi(|Z_{k,i}^{W}|) = 2 \cdot \frac{1}{\sqrt{2\pi}} \int_{|Z_{k,i}^{W}|}^{\infty} \exp\left\{\frac{-\lambda^2}{2}\right\} d\lambda. \tag{13}$$

When $p_{k,i}^{W}$ falls below the threshold $\tau_i^{W} = \alpha_i^{\text{des}}$, i.e., $p_{k,i}^{W} < \tau_i^{W}$, we reject $\mathcal{H}_0$ from (8) and an alarm $\psi_{k,i}^{W} = 1$ is triggered, otherwise $\psi_{k,i}^{W} = 0$. In the absence of attacks, the alarm rate α_i^{W} for an i^{th} sensor should be approximately the same as the desired false alarm rate $\alpha_i^{W} \sim \alpha_i^{\text{des}}$. Computation of α_i^{W} is over the sliding window $T^{\alpha} = (k-\ell^{\alpha}+1, k)$ of length ℓ^{α} by $\alpha_i^{W} = \frac{1}{\ell^{\alpha}} \sum_{j=k-\ell^{\alpha}+1}^{k} \psi_{j,i}^{W}$. Conversely, an attack that affects the residual distribution symmetry, triggering the alarm $\psi_{k,i}^{W}$ more frequently, causing an elevation of alarm rate α_i^{W}. For alarm rates exceeding a user defined alarm rate threshold, i.e. $\alpha_i^{W} > \alpha_i^{\tau}$, the i^{th} sensor is deemed compromised. In the following lemma we provide a proof for bounds of the WSR test variables (10) to satisfy a desired false alarm rate α_i^{des}.

Lemma 1: Given the residual $r_{k,i}$ for an i^{th} sensor over a monitoring window T consisting of ℓ residuals and desired false alarm rate α_i^{des}, an alarm is triggered by the WSR test when $\Omega_{-}^{W} \leq \{W_{k,i}^{\pm}\} \leq \Omega_{+}^{W}$ is not satisfied where

$$\Omega_{\pm}^{W} = \pm|\Phi^{-1}(\alpha_i^{\text{des}}/2)|\sqrt{(\ell^2+\ell)(2\ell+1)/24} + (\ell^2+\ell)/4. \tag{14}$$

Proof: From the Wilcoxon test statistic equaling the sum of ranks in (10), we can rearrange (12) such that $\min(W_{k,i}^{\pm}) = Z_{k,i}^{W_{\text{crit}}} \sqrt{(\ell^2+\ell)(2\ell+1)/24} + (\ell^2+\ell)/4$ where $Z_{k,i}^{W_{\text{crit}}} = \Phi^{-1}(\alpha_i^{\text{des}}/2)$ is the critical value of $Z_{k,i}^{W}$ for $\min(W_{k,i}^{\pm})$ satisfying a desired alarm rate α_i^{des} to not reject (8). The lower bound of $\{W_{k,i}^{-}, W_{k,i}^{+}\}$ must satisfy

$$\begin{aligned} \Omega_{-}^{W} &= \Phi^{-1}(\alpha_i^{\text{des}}/2)\sqrt{(\ell^2+\ell)(2\ell+1)/24} \\ &\quad + (\ell^2+\ell)/4 \leq \min(W_{k,i}^{-}, W_{k,i}^{+}), \end{aligned} \tag{15}$$

to not sound off an alarm $\psi_{k,i}^{W}$. Conversely, we want to show that if the lower bound $\Omega_{-}^{W} \leq \min(W_{k,i}^{\pm})$ in (15) holds then the upper bound Ω_{+}^{W} holds as well. By again manipulating (12) such that we take the maximum $\max(W_{k,i}^{\pm}) = Z_{k,i}^{W_{\text{crit}}} \sqrt{(\ell^2+\ell)(2\ell+1)/24} + (\ell^2+\ell)/4$ where this time $Z_{k,i}^{W_{\text{crit}}} = \Phi^{-1}(1-\alpha_i^{\text{des}}/2)$ is the critical value of $Z_{k,i}^{W}$ for the upper bound $\max(W_{k,i}^{\pm})$ satisfying a desired alarm rate α_i^{des} to not reject (8), the upper bound is written as

$$\begin{aligned} \Omega_{+}^{W} &= \Phi^{-1}(1-\alpha_i^{\text{des}}/2)\sqrt{(\ell^2+\ell)(2\ell+1)/24} \\ &\quad + (\ell^2+\ell)/4 \geq \max(W_{k,i}^{-}, W_{k,i}^{+}), \end{aligned} \tag{16}$$

[1]For window length of smaller size, exact tables need to be used for probability distributions of the Wilcoxon Signed-Rank random variable [14].

to not trigger the alarm $\psi^W_{k,i}$. In the calculation of the critical z-score value from the standard normal distribution $\mathcal{N}(0,1)$ to satisfy a given desired alarm rate α^{des}_i, it is easy to show that $|\Phi^{-1}(\alpha^{\text{des}}_i/2)| = \Phi^{-1}(1-\alpha^{\text{des}}_i/2)$ and $\Phi^{-1}(\alpha^{\text{des}}_i/2) = -|\Phi^{-1}(\alpha^{\text{des}}_i/2)|$ giving the final bounds of $\Omega^W_- \leq (W^{\pm}_{k,i} = \{W^-_{k,i}, W^+_{k,i}\}) \leq \Omega^W_+$ as

$$-|\Phi^{-1}(\alpha^{\text{des}}_i/2)|\sqrt{(\ell^2+\ell)(2\ell+1)/24} + (\ell^2+\ell)/4 \leq W^{\pm}_i \leq |\Phi^{-1}(\alpha^{\text{des}}_i/2)|\sqrt{(\ell^2+\ell)(2\ell+1)/24} + (\ell^2+\ell)/4,$$

satisfying the bounds of $\Omega^W_{\pm}$ in (14). With this we conclude that if $\min(W^{\pm}_{k,i})$ does not satisfy (15) then $\Omega^W_- \leq \{W^-_{k,i}, W^+_{k,i}\} \leq \Omega^W_+$ is not satisfied, triggering alarm $\psi^W_{k,i}$ for a desired false alarm rate α^{des}_i, ending the proof. ■

B. Serial Randomness Monitor

The WSR test alone is not sufficient to test for randomness, since an attacker could manipulate measurements by creating specific patterns to avoid detection on the WSR test. To test further, we need to determine if the sequence of residuals are being received randomly by leveraging the Serial Independence runs (SIR) test [3]. The SIR test examines the number of runs that occur over the sequence, where a "run" is defined as one or more consecutive residuals that are greater or less than the previous value. A random sequence of residuals over a given window length should exhibit a specific expected number of runs: too many or too few number of runs would not satisfy random sequential behavior. A hypothesis test is formed with $\mathcal{H}_0$ for the absence of sensor attacks and $\mathcal{H}_a$ where attacks are present by

$$\mathcal{H}_0\text{: } N_R = \text{E}[N_R], \quad \mathcal{H}_a\text{: } N_R \neq \text{E}[N_R], \tag{17}$$

where N_R is the number of observed runs, to determine whether the number of runs satisfy a randomly behaving sequence. First, we take the difference of residuals at time instances k and $k-1$ over a window T'

$$\boldsymbol{r}'_{T',i} := r'_{k,i} = r_{k,i} - r_{k-1,i}\ ,\ k \in T', \tag{18}$$

where $T' = \{k-\ell+2, \ldots, k\} = T \backslash \{k-\ell+1\}$ is the monitor window T shortened by one by removing the oldest time instance. This in turn gives us $\ell' = \ell - 1$ residual differences.

***Remark** 2:* A residual difference $r'_{k,i} = 0$, $k \in T'$ from (18) is not considered in the test and the size of ℓ' is reduced accordingly, i.e., $\ell' = \ell' - 1$.

From the sequence of residual differences (18), we take the sign of each residual within the window T',

$$\text{sign}(\boldsymbol{r}'_{k,i}),\ k \in T', \tag{19}$$

forming a sequence of ℓ' positive and negative signs. The number of runs N_R are observed over the sequence of ℓ' residual differences. The expected mean and variance of runs [3] are computed by

$$\text{E}[N_R] = \frac{2\ell'-1}{3}, \quad \text{Var}[N_R] = \frac{16\ell'-29}{90}. \tag{20}$$

Assuming large data sets (i.e. window length $\ell \geq 25$) [3], the distribution of N_R converges to a normal distribution as $\ell' \to \infty$ and can be approximated to a zero mean unit variance standard normal distribution $N_R \sim \mathcal{N}(0,1)$. From the number of observed runs N_R and number of residual differences ℓ', we compute the z-score test statistic for Serial Independence from a standard normal distribution

$$Z^S_{k,i} = \frac{N_R - \text{E}[N_R]}{\sqrt{\text{Var}[N_R]}} = \frac{N_R - \big(2\ell'-1\big)/3}{\sqrt{\big(16\ell'-29\big)/90}}. \tag{21}$$

Using the z-score from (21) we compute the p-value of the observed signed residual differences by

$$p^S_{k,i} = \Phi(|Z^S_{k,i}|) = 2 \cdot \frac{1}{\sqrt{2\pi}} \int_{|Z^S_{k,i}|}^{\infty} \exp\left\{\frac{-|\lambda|^2}{2}\right\} d\lambda. \tag{22}$$

When $p^S_{k,i} < \tau^S_i$ is satisfied where $\tau^S_i = \alpha^{\text{des}}_i$ denotes the threshold, we reject the null hypothesis $\mathcal{H}_0$ from (17) and an alarm $\psi^S_{k,i} = 1$ is triggered. In the absence of attacks, the alarm rate α^S_i is approximately the same as the desired false alarm rate $\alpha^S_i \sim \alpha^{\text{des}}_i$. Alarm rate α^S_i over the sliding window T^α is computed by $\alpha^S_i = \frac{1}{\ell^\alpha}\sum_{j=k-\ell^\alpha+1}^{k} \psi^S_{j,i}$. Alarm rates exceeding a user defined alarm rate threshold, i.e. $\alpha^S_i > \alpha^\tau_i$, signifies that the i^{th} sensor is compromised.

***Remark** 3:* A special case of triggering alarm $\psi^S_{k,i} = 1$ is when Remark 2 is satisfied, when two consecutive residuals are equal. Since $r_{k,i} \sim \mathcal{N}(0, \sigma^2_i)$, the probability of having two residuals of the same value is equal to 0.

The following lemma provides a proof for bounds of N_R in the SIR test to satisfy a desired false alarm rate α^{des}_i.

***Lemma** 2:* Given the residual differences $r'_{k,i} = r_{k,i} - r_{k-1,i}$ for an i^{th} sensor over a window T' and desired false alarm rate α^{des}_i, an alarm is triggered by the SIR test when $\Omega^S_- \leq N_R \leq \Omega^S_+$ is not satisfied where

$$\Omega^S_{\pm} = \pm|\Phi^{-1}(\alpha^{\text{des}}_i/2)|\sqrt{(16\ell'-29)/90} + (2\ell'-1)/3. \tag{23}$$

Proof: With an observed number of runs N_R within a window of ℓ' residual differences, we can rearrange (21) such that $N_R = |Z^S_{k,i}|\sqrt{(16\ell'-29)/90} + (2\ell'-1)/3$ where $|Z^S_{k,i}| = |\Phi^{-1}(\alpha^{\text{des}}_i/2)|$, we find the bounds of N_R to not reject (17) for a desired false alarm rate α^{des}_i are

$$-|\Phi^{-1}(\alpha^{\text{des}}_i/2)|\sqrt{(16\ell'-29)/90} + (2\ell'-1)/3 \leq N_R \leq |\Phi^{-1}(\alpha^{\text{des}}_i/2)|\sqrt{(16\ell'-29)/90} + (2\ell'-1)/3. \tag{24}$$

From (24) we can finally obtain the bounds of $\Omega^S_{\pm}$ in (23) for alarm triggering at a desired false alarm rate α^{des}_i. ■

IV. Stealthy Attack Analysis

This section analyzes the advantages of including the proposed randomness monitoring framework into well known boundary/bad-data attack detectors. To this end, we first introduce two well known anomaly (boundary) detectors – Bad-Data [4] and Cumulative Sum [9] detectors – and analyze the effects of stealthy attacks on a system with and without our Randomness Monitor.

A. Boundary Detectors

To show that our framework can easily be augmented with any detector that provides magnitude boundaries, we consider two different boundary detectors found in the CPS security literature. Both boundary detectors discussed in this section leverage the absolute value of the residual (4) for attack detection. Consequently, in the absence of attacks (i.e. $\boldsymbol{\xi}_k = \mathbf{0}$), this leads to $|r_{k,i}|$ following a half-normal distribution [15] defined by

$$\mathrm{E}[|r_{k,i}|] = \sqrt{2/\pi}\sigma_i,\ \mathrm{Var}[|r_{k,i}|] = \sigma_i^2(1-2/\pi). \quad (25)$$

where σ_i^2 was defined as the i^{th} diagonal element in (5).

The first detector that we consider is the *Bad-Data Detector* (BDD) [4], a benchmark attack detector to find anomalies in sensor measurements, alarming when the residual error goes beyond a threshold. Similar to our detection framework in Section III, the BDD can also be tuned for a desired false alarm rate α_i^{des}. Considering the residual $r_{k,i}$ in (4), the BDD procedure for each i^{th} sensor is as follows:

Bad-Data Detector Procedure

$$\textbf{If } |r_{k,i}| > \tau_i^B, \text{ then alarm } \psi_{k,i}^B = 1,\ i \in \mathcal{S}, \quad (26)$$

Assuming the system is without attacks, the tuned threshold τ_i^B for the BDD in (26) with $r_{k,i} \sim \mathcal{N}(0, \sigma_i^2)$ is solved by $\tau_i^B = \sqrt{2}\sigma_i \mathrm{erf}^{-1}(1-\alpha_i^{\text{des}})$ where $\mathrm{erf}^{-1}(\cdot)$ is the *inverse error function*, resulting in $\alpha_i^B \sim \alpha_i^{\text{des}}$.

The second well-known boundary detector that we consider is the *CUmulative SUM* (CUSUM), which has been shown to have tighter bounds on attack detection than the BDD [9]. The CUSUM leverages the absolute value of the residual in the detection procedure and is solved by

CUSUM Detector Procedure

$$\begin{aligned} &\textbf{Initialize } S_{1,i} = 0,\ i \in \mathcal{S}, \\ &S_{k,i} = \max(0, S_{k-1,i} + |r_{k,i}| - b_i), \ \textbf{if } S_{k-1,i} \le \tau_i^C, \\ &S_{k,i} = 0 \text{ and Alarm } \psi_{k,i}^C = 1, \qquad \textbf{if } S_{k-1,i} > \tau_i^C. \end{aligned} \quad (27)$$

The working principle of of this detector is to accumulate the residual sequence in $S_{k,i}$, triggering an alarm $\psi_{k,i}^C = 1$ when the test variable surpasses the threshold τ_i^C. A detailed explanation of how to tune threshold τ_i^C given a bias b_i for a desired false alarm rate α_i^{des} can be found in [9].

B. State Deviation under Worst-case Stealthy Attacks

We consider the reference tracking feedback controller

$$\boldsymbol{u}_k = \boldsymbol{K}\hat{\boldsymbol{x}}_k + \boldsymbol{k}_r \boldsymbol{x}_k^{\text{ref}}, \quad (28)$$

where $\boldsymbol{K} \in \mathbb{R}^{s\times n}$ is the state feedback control gain matrix, $\boldsymbol{k}_r \in \mathbb{R}^{m\times m}$ is a reference gain for output tracking, $\boldsymbol{x}_k^{\text{ref}}$ is the reference state, and $\hat{\boldsymbol{x}}_k$ is the KF state estimate from (2)-(3). Choosing a suitable $\boldsymbol{K}$ such that $(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{K})$ is stable (i.e. $\rho[\boldsymbol{A}+\boldsymbol{B}\boldsymbol{K}] < 1$, where $\rho[\cdot]$ is the spectral radius) and $(\boldsymbol{A}, \boldsymbol{C})$ is assumed stabilizable, the closed-loop system can be written within terms of the KF estimation error as

$$\begin{aligned} \boldsymbol{x}_{k+1} &= (\boldsymbol{A}+\boldsymbol{B}\boldsymbol{K})\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{k}_r\boldsymbol{x}_k^{\text{ref}} - \boldsymbol{B}\boldsymbol{K}\boldsymbol{e}_k + \boldsymbol{\nu}_k, \\ \boldsymbol{e}_{k+1} &= (\boldsymbol{A}-\boldsymbol{L}\boldsymbol{C})\boldsymbol{e}_k - \boldsymbol{L}(\boldsymbol{\xi}_k + \boldsymbol{\eta}_k) + \boldsymbol{\nu}_k. \end{aligned} \quad (29)$$

As an attacker injects signals into the measurements (i.e. $\boldsymbol{\xi} \neq \boldsymbol{0}$), system dynamics are indirectly affected via the interconnected term $\boldsymbol{B}\boldsymbol{K}\boldsymbol{e}_k$ from the estimation error dynamics.

In the remaining of this section we describe the maximum damage that can occur due to worst-case scenario stealthy sensor attacks. We assume the attacker has perfect knowledge of system dynamics, detection procedures, and state estimation. The objective of an attacker is to cause maximum damage to the system state by injecting attack signals $\boldsymbol{\xi}_k$ to measurements while also remaining undetected. With only the BDD implemented, the effects of a worst-case scenario attack while not triggering an alarm can be derived from (4) and (26) with a sustained attack signal

$$\xi_{k,i} = -\boldsymbol{C}_i\boldsymbol{e}_k - \eta_{k,i} + \tau_i^B, \quad (30)$$

causing the residual $|r_{k,i}| = \tau_i^B$ to not trigger the BDD alarm.

Now considering CUSUM as a stand-alone detector, an adversarial wants to avoid attack vectors such that the monitoring test variable exceeds threshold τ_i^C, thereby causing a reset $S_{k,i} = 0$, if $S_{k-1,i} > \tau_i^C$ in (27) by satisfying the CUSUM procedure sequence $S_{k,i} = \max(0, S_{k-1,i} + |\boldsymbol{C}_i\boldsymbol{e}_k + \eta_{k,i} + \xi_{k,i}| - b_i) \le \tau_i^C$ if $S_{k-1,i} \le \tau_i^C$. For maximum effect on state deviation, the attacker saturates the CUSUM test statistic such that $S_{k,i} = \tau_i^C$ to achieve no alarm sequences. Here we define a saturation as follows:

***Definition** 2: Saturation* of a boundary detector is defined as the maximum allowable attack signal to force the residual to, but without exceeding, the detector threshold.

Beginning at a time k, an attacker immediately saturates $S_{k,i}$ with the attack signal,

$$\xi_{k,i} = -\boldsymbol{C}_i\boldsymbol{e}_k - \eta_{k,i} + b_i - S_{k-1,i} + \tau_i^C, \quad (31)$$

followed by

$$\xi_{k,i} = -\boldsymbol{C}_i\boldsymbol{e}_k - \eta_{k,i} + b_i. \quad (32)$$

for all future time instances to hold $S_{k,i}$ at threshold τ_i^C.

With the Randomness Monitor augmented with either BDD or CUSUM, an attacker can no longer hold an attack sequence to one side as described in attacks (30)-(32). Rather, an attacker is forced to create an attack sequence such that $r_{k,i}$ alternates residual signs if it wants to avoid triggering alarms for both the WSR and SIR tests. The most effective attack for maximum state deviation with our augmented framework is to *saturate* the boundary detector as often as possible, while leaving the remaining attack signals with an opposite sign with respect to the saturating attacks to force the residual to be as close as possible to zero.

From the WSR test, given a monitoring window ℓ, the minimum number of *non-saturating* attack signals $\xi_{k,i}$ to not trigger an alarm $\psi_{k,i}^W$ is

$$\gamma_i^\ell = \min_{\ell^j} \left(\sum_{rank=1}^{\ell^j} rank \right) \Bigg| \sum_{rank=1}^{\ell^j} rank > \min(W_i^{\pm}), \quad (33)$$

in which $\ell^j \in \mathcal{L} = (1, \ldots, \ell)$ and $\mathcal{L}$ is the set of all $ranks$ as introduced in Section III-A. From (33), we can then find the maximum number of *saturating* attack signals by $\beta_i^\ell = \ell - \gamma_i^\ell$.

***Proposition** 1:* The maximum allowable saturating attack signal converges to $\lim_{\ell\to\infty} \frac{\beta_i^\ell}{\ell} = 1 - \frac{\sqrt{2}}{2} \approx .293$ for any α_i^{des} as shown by the dashed black line in Fig 3.

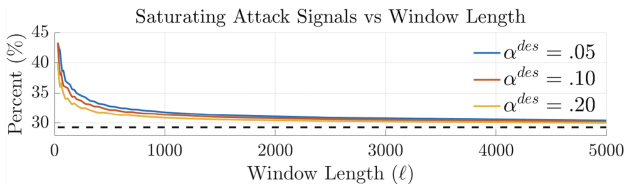

Fig. 3. Allowable percentage of saturating attack signals of given windows lengths for different desired alarm rates α^{des}.

To this point, we have discussed worst-case scenario attack sequences causing saturation of the test variable (in this paper BDD and CUSUM) to maximize the effect of the attack. However, from Remark 3 in Section III-B, a special case to satisfy requirements of the SIR test is when two consecutive residuals of same value triggers an alarm $\psi_{k,i}^S = 1$. To work around this issue, a stealthy attacker with perfect knowledge

of the SIR test can include a small uniformly random number to the attack signal $\xi_{k,i}$ denoted by $\delta_{k,i} \sim \mathcal{U}(0,\epsilon)$ where $\epsilon \in \mathbb{R}^+$ is infinitesimally small and $\mathrm{E}[\delta_{k,i}] = \frac{\epsilon}{2} \approx 0$. Thus, the worst-case scenario with the Randomness Monitor augmented to the BDD follows

$$\begin{cases} \xi_{k,i} = -\boldsymbol{C}_i \boldsymbol{e}_k - \eta_{k,i} + \tau_i^B - \delta_{k,i}, & \textbf{if saturating,} \\ \xi_{k,i} = -\boldsymbol{C}_i \boldsymbol{e}_k - \eta_{k,i} - \delta_{k,i}, & \textbf{if non-saturating,} \end{cases} \tag{34}$$

in order to not trigger an alarm. Similarly, but with the CUSUM detector, an undetectable attack sequence follows

$$\begin{cases} \xi_{k,i} = -S_{k-1,i} - \boldsymbol{C}_i \boldsymbol{e}_k \\ \qquad\quad - \eta_{k,i} + b_i + \tau_i^C - \delta_{k,i}, & \textbf{if saturating,} \\ \xi_{k,i} = -\boldsymbol{C}_i \boldsymbol{e}_k - \eta_{k,i} + b_i - \delta_{k,i}, & \textbf{if non-saturating.} \end{cases} \tag{35}$$

Given the alternating signed sequence of residuals over the monitoring window, the expected value of $r_{k,i}$ under worst-case scenario stealthy attacks is denoted as

$$\begin{cases} \mathrm{E}[r_{k,i}^B] = \tau_i^B(\frac{\beta_i^\ell}{\ell} - \delta_{k,i}) \approx \tau_i^B \frac{\beta_i^\ell}{\ell}, & \textbf{for Bad-Data,} \\ \mathrm{E}[r_{k,i}^C] = \tau_i^C(\frac{\beta_i^\ell}{\ell} - \delta_{k,i}) \approx \tau_i^C \frac{\beta_i^\ell}{\ell}, & \textbf{for CUSUM.} \end{cases} \tag{36}$$

With our framework augmented to the BDD, the expected value of the residual sequence is described as $\mathrm{E}[\boldsymbol{r}_k^B] = (\mathrm{E}[r_{k,1}^B], \ldots, \mathrm{E}[r_{k,s}^B])^T$ and the expectation of the closed-loop system (29) under attack (34) results in

$$\begin{aligned} \mathrm{E}[\boldsymbol{x}_{k+1}] &= (\boldsymbol{A} + \boldsymbol{BK})\mathrm{E}[\boldsymbol{x}_k] - \boldsymbol{BK}\mathrm{E}[\boldsymbol{e}_k], \\ \mathrm{E}[\boldsymbol{e}_{k+1}] &= \boldsymbol{A}\mathrm{E}[\boldsymbol{e}_k] - \boldsymbol{L}\mathrm{E}[\boldsymbol{r}_k^B]. \end{aligned} \tag{37}$$

Note, in (37), the reference signal term $\boldsymbol{Bk}_r \boldsymbol{x}_k^{\text{ref}}$ from (29) has been removed as we are interested in the expected state deviation under an attack. It is clear that if $\rho[\boldsymbol{A}] > 1$ and $\mathrm{E}[\boldsymbol{r}_k^B] \neq \boldsymbol{0}$ then the expectation of the estimation error $\mathrm{E}[\boldsymbol{e}_k]$ for destabilized states diverge to infinity as $k \to \infty$ (depending on algebraic properties of $\boldsymbol{A}$), indirectly causing these states within $\mathrm{E}[\boldsymbol{x}_k]$ to also diverge unbounded.

Lemma 3: Considering a closed-loop system from (1) and (37), where $\rho[\boldsymbol{A}] < 1$ and attack sequence in (34), the limit for expected state divergence $\lim_{k\to\infty} \mathrm{E}[\boldsymbol{x}_k] = \Delta^B$ is

$$\Delta^B = (\boldsymbol{I} - \boldsymbol{A} - \boldsymbol{BK})^{-1}\boldsymbol{BK}(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{L}\mathrm{E}[\boldsymbol{r}_k^B]. \tag{38}$$

Proof: Assuming both $\rho[\boldsymbol{A}] < 1$ and $\rho[\boldsymbol{A} + \boldsymbol{BK}] < 1$ are satisfied, signifying the invertibility of $(\boldsymbol{I} - \boldsymbol{A})$ and $(\boldsymbol{I} - \boldsymbol{A} - \boldsymbol{BK})$ in (38), an expected equilibrium is reached as $k \to \infty$ by $\mathrm{E}[\boldsymbol{x}_\infty] = (\boldsymbol{I} - \boldsymbol{A} - \boldsymbol{BK})^{-1}\boldsymbol{BK}(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{L}\mathrm{E}[\boldsymbol{r}_k^B]$ and $\mathrm{E}[\boldsymbol{e}_\infty] = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{L}\mathrm{E}[\boldsymbol{r}_k^B]$ such that the evolution of (37) with the expected differences $\mathrm{E}[\boldsymbol{x}_k] - \mathrm{E}[\boldsymbol{x}_\infty]$ and $\mathrm{E}[\boldsymbol{e}_k] - \mathrm{E}[\boldsymbol{e}_\infty]$ is described by

$$\begin{aligned} \mathrm{E}[\boldsymbol{x}_{k+1}] - \mathrm{E}[\boldsymbol{x}_\infty] &= (\boldsymbol{A} + \boldsymbol{BK})(\mathrm{E}[\boldsymbol{x}_k] - \mathrm{E}[\boldsymbol{x}_\infty]) \\ &\quad - \boldsymbol{BK}(\mathrm{E}[\boldsymbol{e}_k] - \mathrm{E}[\boldsymbol{e}_\infty]), \\ \mathrm{E}[\boldsymbol{e}_{k+1}] - \mathrm{E}[\boldsymbol{e}_\infty] &= \boldsymbol{A}\mathrm{E}[\boldsymbol{e}_k] - \mathrm{E}[\boldsymbol{e}_\infty], \end{aligned} \tag{39}$$

are asymptotically stable i.e., $\lim_{k\to\infty}(\mathrm{E}[\boldsymbol{x}_{k+1}] - \mathrm{E}[\boldsymbol{x}_\infty]) = \boldsymbol{0}$ and $\lim_{k\to\infty}(\mathrm{E}[\boldsymbol{e}_{k+1}] - \mathrm{E}[\boldsymbol{e}_\infty]) = \boldsymbol{0}$, therefore concluding the proof. ∎

Similarly, with the Randomness Monitor augmented to CUSUM, the expected closed-loop system evolution under attack sequence (35) is described by

$$\begin{aligned} \mathrm{E}[\boldsymbol{x}_{k+1}] &= (\boldsymbol{A} + \boldsymbol{BK})\mathrm{E}[\boldsymbol{x}_k] - \boldsymbol{BK}\mathrm{E}[\boldsymbol{e}_k], \\ \mathrm{E}[\boldsymbol{e}_{k+1}] &= \boldsymbol{A}\mathrm{E}[\boldsymbol{e}_k] - \boldsymbol{L}\mathrm{E}[\boldsymbol{r}_k^C]. \end{aligned} \tag{40}$$

where $\mathrm{E}[\boldsymbol{r}_k^C] = (\mathrm{E}[r_{k,1}^C], \ldots, \mathrm{E}[r_{k,s}^C])^T$ is the expected value of the residual sequence vector for CUSUM in (36).

Lemma 4: Considering a closed-loop system from (1) and (40), where $\rho[\boldsymbol{A}] < 1$ and attack sequence in (35), the limit for expected state divergence $\lim_{k\to\infty} \mathrm{E}[\boldsymbol{x}_k] = \Delta^C$ is

$$\Delta^C = (\boldsymbol{I} - \boldsymbol{A} - \boldsymbol{BK})^{-1}\boldsymbol{BK}(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{L}\mathrm{E}[\boldsymbol{r}_k^C]. \tag{41}$$

Proof: The proof is omitted here due to space constraints but follows the proof for Lemma 3. ∎

V. Results

The proposed Randomness Monitor framework was validated in simulation and experiments and compared to state-of-the-art detection techniques introduced in Section IV-B. The case study presented in this paper is an autonomous waypoint navigation of a skid-steering differential-drive UGV with the following linearized model [16]

$$\begin{aligned} \dot{v} &= \frac{1}{m}(F_l + F_r - B_r v), \\ \dot{\omega} &= \frac{1}{I_z}\Big(\frac{w}{2}(F_l - F_r) - B_l\omega\Big), \ \dot{\theta} = \omega, \end{aligned} \tag{42}$$

where v is the velocity, θ is the vehicle's heading angle, and ω its angular velocity, forming the state vector $\boldsymbol{x} = [v, \theta, \omega]^T$. F_l and F_r describe the left and right input forces from the wheels, w is the vehicle width, while B_r and B_l are resistances due to the wheels rolling and turning. The continuous-time model (42) is discretized with a sampling rate $t_s = 0.05$ to satisfy the system model described in (1).

In both simulation and experiment we perform two different attack sequences: *Attack #1* where a stealthy attack sequence concentrates the residual distribution with a non-zero mean and smaller variance whereas *Attack #2* creates a signed pattern sequence {+, +, +, -} of residual differences $r'_{k,i}$. Both attacks are chosen to not increase the boundary detector alarm rate.

A. Simulations

Considering the UGV system model (42) in our case study, we show the effect of stealthy attacks on the velocity sensor on state x_1 with a monitoring window length $\ell = 100$. Table I gives the resulting alarm rates when our framework is augmented to boundary detectors (BDD and CUSUM) with all detectors tuned for desired false alarm rates $\alpha^{\text{des}} \in \{.05, .20\}$ and in separate simulations we show the alarm rate for *No Attack*, *Attack #1*, and *Attack #2*. As expected, with no attacks present, all alarm rates converge approximately to the desired false alarm rate α_1^{des}. Under *Attack #1*, alarm rates for only the WSR increase to higher values and similarly the *Attack #2* pattern gives an increased alarm rate to only the SIR test. We should note that the window length ℓ results in different behaviors: short window lengths result in faster responses, while longer window lengths react slower but are able to detect more hidden attacks exhibiting non-random behavior than a monitor with a short window length. Fig. 4 demonstrates attacks on the velocity sensor where our detectors are tuned for $\alpha_1^{\text{des}} = 0.10$ and compared with the CUSUM boundary detector. *Attack #1* occurs between $(50, 125)$s triggering the WSR test, *Attack #2* between $(175, 250)$s triggering the SIR test, and from 300s a third attack satisfying bounds for both randomness tests but

violating the CUSUM test is presented. Velocity is reduced as expected according to (29) while experiencing the effects of each attack.

TABLE I
SIMULATED ALARM RATES

State x_1 with $\ell = 100$		Randomness Monitor		Boundary Detectors	
$b_1 = 1.10$ *for CUSUM*		WSR	SIR	BDD	CUSUM
$\alpha^{des} = .20$	α_1 (No attack)	0.1809	0.2133	0.1941	0.1771
	α_1 (Attack 1)	0.9945	0.2089	0.1555	0.1569
	α_1 (Attack 2)	0.1874	1.0000	0.1800	0.2000
$\alpha^{des} = .05$	α_1 (No attack)	0.0392	0.0492	0.0519	0.0340
	α_1 (Attack 1)	0.9982	0.0513	0.0393	0.0361
	α_1 (Attack 2)	0.0377	1.0000	0.0500	0.0500

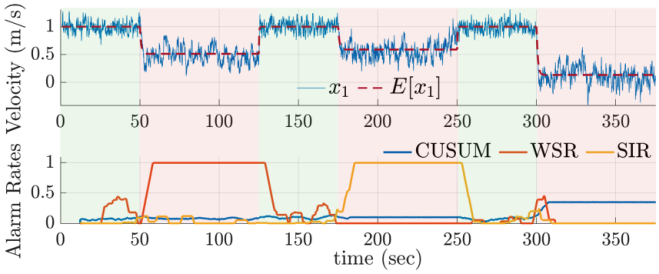

Fig. 4. State deviation under various attacks and alarm rates over a moving window of the past 100 time steps.

B. Experiments

In this section we present a case study for a UGV performing way-point navigation under stealthy sensor attacks. For our case, the UGV travels to a series of goal positions while avoiding a restricted area with a desired cruise velocity $v^{\text{ref}} = 0.15$m/s while experiencing the same class of attacks as in Section V-A. This time the IMU sensor that measures angle θ is spoofed while our Randomness Monitor is augmented with the BDD. Fig. 5 shows the UGV position while traveling to the four goal points. For both attacks the vehicle enters the restricted area (marked by red tape) while the boundary detector (BDD) does not see the attack in each case. The alarm rate for the WSR test increases for the case under *Attack #1* (solid line) and the SIR test alarm rate increases during the case for *Attack #2* (dashed line), as expected.

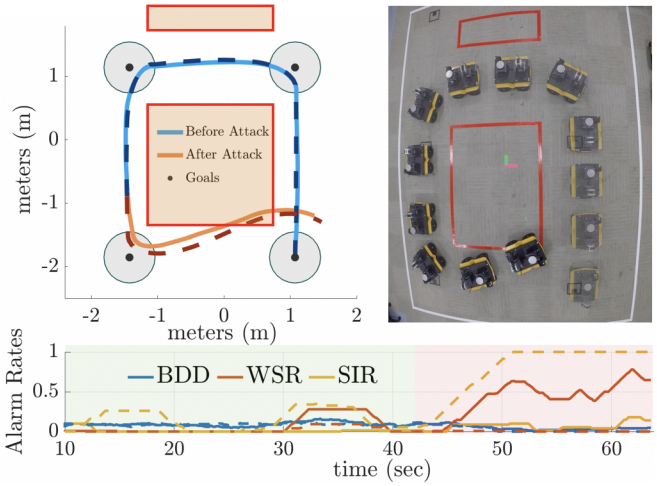

Fig. 5. UGV position under *Attack #1* (solid line) and *Attack #2* (dashed line). The bottom graph displays the resulting alarm rates.

VI. CONCLUSIONS & FUTURE WORK

In this paper we have proposed a monitoring framework to find cyber-attacks that present non-random behavior with the intention to hijack a system from a desired state. Our framework leverages the Wilcoxon Signed-Rank test and Serial Independence Runs test over a sliding monitor window to detect stealthy attacks when augmented to state-of-the-art boundary detectors. Among the key results of this work we provide: bounds for desired false alarm rate for each test which are leveraged to detect attacks, bounds on state deviation under worst case attack scenario, demonstrating that the proposed framework outperform detectors that solely use boundary tests. The proposed approach was validated through simulations and experiments on UGV case studies.

In our future work we plan to extend the current work to remove this dependency from the monitoring window and plan to leverage our approach in systems with redundant sensors to remove the compromised sensors and build attack resilient controllers similar to our previous work in [6].

ACKNOWLEDGMENTS

This work is based on research sponsored by ONR under agreement number N000141712012, and NSF under grant #1816591.

REFERENCES

[1] J. Bhatti and T. E. Humphreys, "Hostile control of ships via false gps signals: Demonstration and detection," *Navigation*, vol. 64, no. 1, pp. 51–66, 2017.
[2] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
[3] C. Cammarota, "The difference-sign runs length distribution in testing for serial independence," *Journal of Applied Statistics*, vol. 38, no. 5, pp. 1033–1043, 2011.
[4] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *2010 IEEE 49th Conference on Decision and Control*, pp. 5967–5972.
[5] C. Bai and V. Gupta, "On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds," in *2014 American Control Conference*, June 2014, pp. 3029–3034.
[6] N. Bezzo, J. Weimer, M. Pajic, O. Sokolsky, G. J. Pappas, and I. Lee, "Attack resilient state estimation for autonomous robotic systems," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2014, pp. 3692–3698.
[7] C. Kwon, W. Liu, and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," in *2013 American Control Conference*, June 2013, pp. 3344–3349.
[8] C. Kwon, S. Yantek, and I. Hwang, "Real-time safety assessment of unmanned aircraft systems against stealthy cyber attacks," *Journal of Aerospace Information Systems*, vol. 13, no. 1, pp. 27–45, 2016.
[9] C. Murguia and J. Ruths, "Characterization of a cusum model-based sensor attack detector," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec 2016, pp. 1303–1309.
[10] C. Murguia and J. Ruths, "On model-based detectors for linear time-invariant stochastic systems under sensor attacks," *IET Control Theory Applications*, vol. 13, no. 8, pp. 1051–1061, 2019.
[11] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding sensor outputs for injection attacks detection," in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 5776–5781.
[12] T. Severson, *et al.*, "Trust-based framework for resilience to sensor-targeted attacks in cyber-physical systems," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 6499–6505.
[13] M. Pajic, J. Weimer, N. Bezzo, O. Sokolsky, G. J. Pappas, and I. Lee, "Design and implementation of attack-resilient cyberphysical systems: With a focus on attack-resilient state estimators," *IEEE Control Systems Magazine*, vol. 37, no. 2, pp. 66–81, April 2017.
[14] S. Siegel, *Nonparametric statistics for the behavioral sciences*. McGraw-Hill New York, 1956.
[15] S. M. Ross, *Introduction to Probability Models, Ninth Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.
[16] J. J. Nutaro, *Building software for simulation: theory and algorithms, with applications in C++*. John Wiley & Sons, 2011.