Low Latency Attack Detection with Dynamic Watermarking for Grid-Connected Photovoltaic Systems

Yaze Li*, *Student Member, IEEE,* Nan Lin, *Student Member, IEEE,*
Jingxian Wu, *Senior Member, IEEE*, Yanjun Pan, *Member, IEEE*, and Yue Zhao, *Senior Member, IEEE*
Department of Electrical Engineering and Computer Science, University of Arkansas, Fayetteville, AR 72701.

***Abstract*—This paper proposes an active low latency attack detection algorithm to improve the cybersecurity of grid-connected photovoltaic (PV) systems. The algorithm is developed by using a hybrid model- and data-driven approach, where the test statistics are formulated by using both the physical model of the PV farm and historical measurements. Unlike most previous detection algorithms that mainly focus on detection accuracy, the proposed algorithm aims at minimizing detection delay while ensuring detection accuracy. The low latency detection algorithm is developed by designing a generalized cumulative sum (CUSUM) detector with a dynamic watermark, which can detect cyberattacks even if the adversary has full knowledge of the system model. To evaluate the performance of the low latency detection algorithm, we propose to measure the stealthiness of cyberattacks by using the Kullback-Leibler (KL) divergence between the pre- and post-attack distributions of the test statistics. Simulation results demonstrate that the proposed algorithm can accurately detect cyberattacks with minimum delays.**

***Index Terms*—Cybersecurity, photovoltaic (PV), dynamic watermarking, cumulative sum (CUSUM)**

I. Introduction

Electricity generated from renewable energy sources (RES) surpassed coal in the US for the first time in 2022 [1]. The rapidly growing grid integration of RES is mainly driven by the growth in solar and wind, which accounts for 14% US domestic electricity generation in 2022 [1]. High renewable penetration relies on interoperable distributed energy resource (DER) grid-support functions with complex control and communication capabilities. The addition of these complex control and communication capabilities increases the vulnerability of the RES, and make them prone to cyberattacks [2]–[5]. Cyberattacks can disrupt normal grid operations by causing system instabilities such as line overloads, frequency and/or voltage violations, reverse power flow, and voltage collapse, especially during heavy load conditions [6], [7]. This necessitates the development of new cybersecurity technologies that can detect and/or mitigate the negative impacts of cyberattacks.

Many existing studies on the cybersecurity of energy systems focus on grid operations by using measurements from the supervisory control and data acquisition (SCADA) systems, the remote terminal units (RTUs), and/or the underlying communication network of the grid [8], [9]. These measurements are important indicators for grid operations, but they are insufficient given that attacks can be launched against local measurements from sensors and actuators of RES, or the local control policies for RES operations.

The rapid advancement of machine learning (ML) during the past decade has driven the development of ML-based cyberattack detection methods. Most of the ML-based methods are data-driven, and they do not require physical models of the system. Various data-driven ML algorithms, such as one-class support vector machines (OCSVMs), random forests (RFs), and principal component analysis (PCA) were applied to multiple sources of time-series data for distributed anomaly detection on a single solar panel [10]. Deep neural networks (DNN) with long short-term memory (LSTM) were applied to detect data integrity attacks by using the Northeast Solar Energy Research Center (NSERC) PV farm dataset [11]. In [12], a bidirectional LSTM (Bi-LSTM) framework is used to detect false data injection (FDI) attacks on a modified IEEE 14-bus system integrated with RES. An autoencoder is later combined with the Bi-LSTM model in [13] to achieve both cyberattck detection and system state forecasting. Raw data collected from micro PMU (μPMU) were used for the detection of cyberattacks on photovoltaic (PV) farms by using data-driven methods such as decision tree (DT) and K-nearest neighbor (KNN) [14]. Most ML approaches require a large amount of data during the offline training stage, and sometimes it might be difficult to obtain a sufficient amount of training data from cyber-physical systems (CPS).

Most ML-based anomaly detection methods are purely data driven, and they do not utilize the physical model of the underlying system. Neural networks such as CNN use convolution kernels, and LSTM captures the dependencies between historical data, achieving higher detection accuracy [15]. The data-driven approach requires a large amount training data from historical measurements.

In contrast to pure data-driven methods, model-based methods utilize the underlying physical models of CPS to monitor system operations. The knowledge of the physical model can help improve detection accuracy and reduce the amount of training data. For example, measurement results can be compared to state estimations in a smart grid, and the residues can then be used for anomaly detection [16], [17]. The detection can be performed by using either a single measurement or a sequence of historical measurements, such as the windowed

The work of was supported in part by the U.S. National Science Foundation (NSF) under Award Numbers ECCS-1711087 and NSF Center for Infrastructure Trustworthiness in Energy Systems (CITES).

χ^2 detector [18].

All these detection methods can be classified as passive methods. One of the limitations of the passive methods is that they might not be able to detect cyberattacks designed by using full knowledge of the system model, as the adversary can use the knowledge of the physical model to match the attacked data with state estimation results.

Dynamic watermarking is an active defense method that adds a small random signal, i.e., "watermark", to the input of the controller [19]. The power of the random signal is small such that it does not disturb normal system operations, and the detector can utilize the statistical distributions of the watermark signal to test the operation conditions of the CPS. Dynamic watermarking was first proposed to improve the performance of χ^2 detector [19], [20]. However, it is unable to detect attacks with post-attack distributions fitting historical measurements, such as the replay attack. This problem can be solved by using two actuator tests with respect to the covariance of the residuals and the correlation between the residuals and watermarks [21]. The two-test dynamic watermarking scheme is used as the active defense method for the automatic generation control (AGC) of a power system, and to detect attacks applied to voltage and current measurements of a grid-connected PV system [22]. The two-test dynamic watermarking algorithm is later extended to general linear time invariant (LTI) systems with a single statistical test based on the Wishart distribution [23], and to linear time varying systems and nonlinear systems in [24].

Even though well-known model-driven and data-driven detection methods exist in power grids, their applications in PV system security are still in its early stages due to the recent rise of the topic of PV system cybersecurity. Most research focus on fault diagnosis for PV systems with no little or no attention to detection delays [25]–[27]. Detection delay is critical to the cybersecurity of energy systems as a shorter delay means a timely response that can minimize the negative impacts of the attacks. There is a fundamental tradeoff between detection delay and detection accuracy [16]. A lower detection delay might be achieved at the cost of detection accuracy, and vice versa. Quickest change detection (QCD) aims at minimizing the detection delay subject to a constraint on satisfactory detection accuracy. QCD is usually implemented by means of sequential analysis such as the sequential probability ratio test (SPRT) [28], the cumulative sum (CUSUM) [29], [30], generalized likelihood ratio (GLR) testing [31], etc. Most algorithms require perfect knowledge of the post-change distribution [32], which is usually difficult, if not impossible, to obtain [33]. A sequential fault detection scheme based on the generalized local likelihood ratio (GLLR) test is used to achieve quickest fault detection in PV systems [34].

The objective of this paper is to develop a low-latency attack detection algorithm for grid-connected PV systems to minimize the detection delay under the constraint of an upper bound of false alarm rate. Since the key component of the PV system is the inverter under classical closed-loop control, we apply the dynamic watermarking algorithm to achieve active defense, enabling the detector to detect cyber attacks originating from attackers who may have the knowledge of the PV inverter structure. The proposed algorithm has four main innovations. First, the algorithm is designed by using a hybrid model- and data-driven approach. We first construct a state-space model for a grid-connected PV farm, the knowledge of which is used to estimate and predict the state information, such as current and voltage, by using a Kalman filter. Key parameters of the filter are estimated and updated by using data collected from the system. Second, the algorithm performs active detection of cyberattacks by using a two-test dynamic watermarking scheme. The statistical tests of the dynamic watermarks are formulated by analyzing the statistical properties of the residuals from state estimation and measurements. Third, unlike existing methods that focus mainly on detection accuracy, the algorithm is developed to minimize the average detection delay (ADD), subject to an upper bound on the probability of false alarm (PFA). The low latency detection algorithm is designed by using a modified CUSUM algorithm that incorporates dynamic watermarks. Fourth, we propose to measure the stealthiness of various cyberattacks by using the Kullback-Leibler (KL) divergence between the pre- and post-attack distributions of the test statistics. The KL divergence provides a quantitative measure on the tradeoff between the stealthiness and the power of a given cyberattack. The KL divergences of several attacks, such as the FDI attack, replay attack, and destabilization attacks are analyzed and evaluated.

The rest of this paper is organized as follows. The modeling of the grid-connected PV farm (including both physical and state space models) and various cyberattacks are introduced in Section II. Section III develops the details of statistical tests with dynamic watermarking by using state estimation results with Kalman filter. The low latency detection algorithm with dynamic watermarking is proposed in Section IV, where the metric KL divergence is introduced to measure the stealthiness of attacks. Simulation results are given in V, and the paper is concluded in Section VI.

II. System Models

This section describes the model, control and dynamics of a grid-connected PV farm. Various cyberattack models that can compromise PV farm operations are also introduced in this section.

A. Modeling Grid-Connected PV Farm

Fig. 1 illustrates the schematic of a typical photovoltaic (PV) inverter. The precision of the amplitude and frequency of the voltage at the output of the PV inverter is critical for proper system operations. The ideal direct current (DC) voltage source at the output of the boost converter is denoted as $V_{\rm DC}$. Denote the phase voltage magnitude connected to the grid as $V_{\rm G}$, with its DQ frame represented as $V_{\rm DG}$ and $V_{\rm QG}$, respectively. The three-phase output current of the inverter is denoted as I_a, I_b, and I_c, respectively. The corresponding DQ frame representation of the three-phase current is $I_{\rm D}$ and $I_{\rm Q}$, respectively. The operation of the PV inverter controller is performed by controlling signals $V_{\rm DG}$, $V_{\rm QG}$, $I_{\rm D}$, and $I_{\rm Q}$, in conjunction with the DQ frame of the three-phase reference

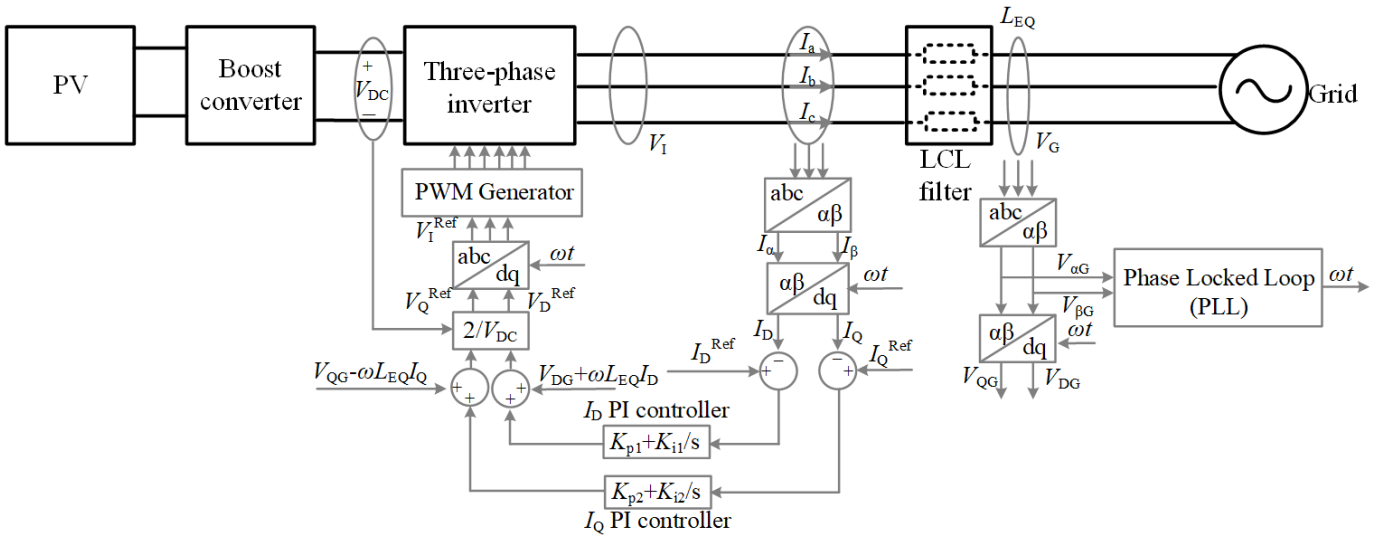

Fig. 1: Diagram of PV inverter.

currents, $I_{\mathrm{D}}^{\mathrm{Ref}}$ and $I_{\mathrm{Q}}^{\mathrm{Ref}}$. The values of these reference currents are established based on the intended output voltage and power to be supplied to the grid.

Denote the phase voltage magnitude of the inverter as V_{I}, and the equivalent inductance between the inverter and the grid as L_{EQ}, which includes the inductance of the LCL filter and the transmission line. Based on the input signals directed towards the PV inverter controller, the computation of the reference DQ frames of V_{I} can be performed as follows,

$$V_{\mathrm{DI}}^{\mathrm{Ref}} = \frac{2}{V_{\mathrm{DC}}}\left[K_{p1}\left(I_{\mathrm{D}}^{\mathrm{Ref}} - I_{\mathrm{D}}\right) + K_{i1}\int_0^t \left(I_{\mathrm{D}}^{\mathrm{Ref}} - I_{\mathrm{D}}\right) d\tau \right. \tag{1a}$$

$$\left. + V_{\mathrm{DG}} + \omega L_{\mathrm{EQ}} I_{\mathrm{Q}}\right],$$

$$V_{\mathrm{QI}}^{\mathrm{Ref}} = \frac{2}{V_{\mathrm{DC}}}\left[K_{p2}\left(I_{\mathrm{Q}}^{\mathrm{Ref}} - I_{\mathrm{Q}}\right) + K_{i2}\int_0^t \left(I_{\mathrm{Q}}^{\mathrm{Ref}} - I_{\mathrm{Q}}\right) d\tau \right. \tag{1b}$$

$$\left. + V_{\mathrm{QG}} - \omega L_{\mathrm{EQ}} I_{\mathrm{D}}\right],$$

where K_{p1}, K_{p2}, K_{i1} and K_{i2} denote the proportional and integral parameters, tuned in accordance with the desired static and dynamic performance standards for the output voltage. The reference voltage $V_{\mathrm{I}}^{\mathrm{Ref}}$ is derived through an inverse DQ transformation originating from its reference DQ frame. Subsequently, this transformed reference voltage is supplied as input to the PWM generator.

B. State-Space model

The PV farm model can be conceptually represented as a multi-input multi-output partially observed system. Based on the grid-connected PV farm model, denote the system state vector as $\mathbf{x} \in \mathcal{R}^n$, the control system input vector as $\mathbf{u} \in \mathcal{R}^m$, and the output (or observation) vector as $\mathbf{y} \in \mathcal{R}^p$, which are defined as follows

$$\mathbf{x} = [I_{\mathrm{D}}, I_{\mathrm{Q}}, V_{\mathrm{DG}}, V_{\mathrm{QG}}]^T, n = 4, \tag{2a}$$

$$\mathbf{u} = [V_{\mathrm{DI}}, V_{\mathrm{QI}}]^T, \quad m = 2, \tag{2b}$$

$$\mathbf{y} = [\omega, |V_{\mathrm{G}}|]^T, \quad p = 2, \tag{2c}$$

where V_{DI} and V_{QI} represent the DQ frames of V_{I}, the voltage at the output of the three-phase inverter. The symbols w and $|V_{\mathrm{G}}|$ denote the frequency and magnitude of the output voltage interlinked with the grid, respectively.

To facilitate the design of the low latency attack detector, the dynamics of the PV farm can be approximated by utilizing the subsequent linearized differential and algebraic equations (DAEs) as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{w}, \tag{3a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{n}, \tag{3b}$$

where $\mathbf{A} \in \mathcal{R}^{n\times n}$, $\mathbf{B} \in \mathcal{R}^{n\times m}$, and $\mathbf{C} \in \mathcal{R}^{p\times n}$ are the state matrix, control matrix, and output matrix, respectively. In addition, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_w})$ and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_n})$ correspond to zero-mean Gaussian-distributed process noise and measurement noise, respectively. It should be noted that the linearized DAE in (3) is only used for the design of the low latency attack detector. The data used in this paper are still generated by using the actual non-linear system as shown in Fig. 1.

The small-signal dynamics around a given operating point can be obtained by using the DAEs. Let $\Delta\mathbf{x}$, $\Delta\mathbf{u}$ and $\Delta\mathbf{y}$ denote the marginal variation from the equilibrium state. The DAEs that represent the small-signal dynamics can then be formulated as follows

$$\Delta\dot{\mathbf{x}} = \mathbf{A}\Delta\mathbf{x} + \mathbf{B}\Delta\mathbf{u} + \mathbf{w}, \tag{4a}$$

$$\Delta\mathbf{y} = \mathbf{C}\Delta\mathbf{x} + \mathbf{n}. \tag{4b}$$

The control policy of the entire PV farm can be abstracted into a nonlinear vector function $\mathbf{h}(\cdot)$ as

$$\Delta\mathbf{u} = \mathbf{h}(\Delta\mathbf{y}). \tag{5}$$

The continuous-time state-space model in (4) can be discretized into discrete-time as

$$\Delta\mathbf{x}[t+1] - \Delta\mathbf{x}[t] = \mathbf{A}\Delta\mathbf{x}[t] + \mathbf{B}\Delta\mathbf{u}[t] + \mathbf{w}[t+1], \quad (6a)$$
$$\Delta\mathbf{y}[t+1] = \mathbf{C}\Delta\mathbf{x}[t+1] + \mathbf{n}[t+1]. \quad (6b)$$

Similarly, the discrete form of (5) is:

$$\Delta\mathbf{u}[t] = \mathbf{h}(\mathbf{y}^t) \quad (7)$$

where $\mathbf{y}^t := \{\mathbf{y}[0], \mathbf{y}[1], \ldots, \mathbf{y}[t]\}$ represents past measurements collected by the sensors in the context of the proportional-integral (PI) control.

It is important to note that the values of variables within the discrete-time difference equations typically diverge from their counterparts in the continuous-time differential equations, and these values are contingent upon the sampling rate. To simplify notation, the same notations are retained, and the small deviation notation Δ is omitted for the remainder of this work. The DAEs presented in (6) can also be reformulated as

$$\mathbf{x}[t+1] = \mathbf{A_d}\mathbf{x}[t] + \mathbf{B}\mathbf{u}[t] + \mathbf{w}[t+1], \quad (8a)$$
$$\mathbf{y}[t+1] = \mathbf{C}\mathbf{x}[t+1] + \mathbf{n}[t+1], \quad (8b)$$

where $\mathbf{A_d} = \mathbf{A} + \mathbf{I}_n$. The state transition matrices for the discrete-time model can be derived from the continuous-time model, provided that the control model is known. Alternatively, these matrices can also be inferred from practical measurements, even in the absence of knowledge regarding the control model.

C. Attack models

Suppose the system is attacked at the moment τ, and assume that the attacker has the knowledge of the control system, including the parameters $\mathbf{A_d}, \mathbf{B}, \mathbf{C}$, the control policy $\mathbf{h}(\cdot)$ and all historical measurements $\mathbf{z}^t$. This is a very generous assumption to assume the worst possible attacks. In case the attacker has partial knowledge of the above parameters and/or control policy, the attack efficiency will be lower and it will be easier to detect.

The following cyberattacks are examined within the scope of this paper.

1) **FDI attack**. The measurement vector $\mathbf{y}[t]$ is injected with a deterministic attack vector $\mathbf{a}[t] \in \mathcal{R}^p$ or a noise vector $\mathbf{a}[t] \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma_a})$ as

$$\mathbf{z}[t] = \begin{cases} \mathbf{y}[t], & t < \tau, \\ \mathbf{y}[t] + \mathbf{a}[t], & t \geq \tau. \end{cases} \quad (9)$$

2) **Replay attack**. The measurement vector $\mathbf{y}[t]$ is replaced by historical data from l moments ago with $l < \tau$ as,

$$\mathbf{z}[t] = \begin{cases} \mathbf{y}[t], & t < \tau, \\ \mathbf{y}[t-l], & t \geq \tau. \end{cases} \quad (10)$$

3) **Destabilization attack**. The control input $\mathbf{u}[t]$ is injected with a scaled controller input as

$$\mathbf{u}_a[t] = \mathbf{u}[t] + \mathbf{A_p}\mathbf{x}[t], \quad t \geq \tau, \quad (11)$$

where $\mathbf{A_p} \in \mathcal{R}^{m \times n}$ is the scaling parameter for the attack. With the compromised control input $\mathbf{u}_a[t]$, the state transition in (8) becomes

$$\mathbf{x}[t+1] = (\mathbf{A_d} + \mathbf{B}\mathbf{A_p})\mathbf{x}[t] + \mathbf{B}\mathbf{u}[t] + \mathbf{w}[t+1]. \quad (12)$$

The instability of the system arises when the elements within matrix $\mathbf{A_p}$ are selected in a manner that satisfies the condition $||\mathbf{A_d} + \mathbf{B}\mathbf{A_p}|| \geq 1$ [35]. This instability holds regardless of the specifics of the control vector, resulting in an inevitable escalation or attenuation of the state vector.

III. Active Attack Detection with Dynamic Watermarking

This section outlines the proposed active attack detection method with dynamic watermarking. The active detection method is motivated by the fact that conventional passive detection methods might not be able to detect attacks designed with full or partial knowledge of the power system.

For example, if an attacker has knowledge of the covariance matrix of the process noise $\mathbf{\Sigma_w}$, then the attacker can replace the true sensor measurements $\mathbf{y}[t]$ with false measurements $\mathbf{y}'[t]$ generated by tracking the following falsified system model

$$\mathbf{x}'[t+1] = \mathbf{A_d}\mathbf{x}'[t] + \mathbf{B}\mathbf{u}[t] + \mathbf{w}'[t+1], \quad (13a)$$
$$\mathbf{y}'[t+1] = \mathbf{C}\mathbf{x}'[t+1] + \mathbf{n}[t+1], \quad (13b)$$

where $\mathbf{x}'[t]$ and $\mathbf{y}'[t]$ are the state vector and measurement vector of the false system, and $\mathbf{w}'[t]$ are artificially generated i.i.d zero-mean noise with covariance matrix $\mathbf{\Sigma_w}$. Conventional passive detection methods are ineffective in identifying this kind of attack, because all state and measurement vectors follow the dynamics of the physical model. Nevertheless, an active defense strategy can be employed by introducing concealed signals, i.e. "dynamic watermarks", which remain undisclosed to both the system and the attacker. As a result, true measurements will exhibit correlation with the dynamic watermark, and such correlation disappears with falsified measurements [21].

The dynamic watermarking is implemented in the form of a random signal $\mathbf{e}[t] \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_e})$, and they are identically and independently distributed in time. The dynamic watermark signal is applied to the control input as,

$$\mathbf{u}[t] = \mathbf{h}(\mathbf{z}^t) + \mathbf{e}[t], \quad (14)$$

where $\mathbf{z}^t$ is the compromised observation vector after attack, and $\mathbf{z}^t = \mathbf{y}^t$ if there is no attack.

With the watermark signal, the system evolves as

$$\mathbf{x}[t+1] = \mathbf{A_d}\mathbf{x}[t] + \mathbf{B}\mathbf{h}(\mathbf{z}^t) + \mathbf{B}\mathbf{e}[t] + \mathbf{w}[t+1], \quad (15a)$$
$$\mathbf{y}[t+1] = \mathbf{C}\mathbf{x}[t+1] + \mathbf{n}[t+1]. \quad (15b)$$

It is shown in [21] that incorporating a dynamic watermark signal into the control input can serve to uncover any illicit manipulation of the signals via the application of two distinct statistical tests.

The development of low latency detection through dynamic watermarking requires estimation and tracking of the state vector of the PV farm. The state estimation is performed by using the Kalman filter as shown in the next subsection.

A. Kalman Filter

Denote $\hat{\mathbf{x}}_{a|b}$ as the estimation of $\mathbf{x}$ at the moment a given observations up to and including moment b.

The prior state estimation and the prior estimation covariance matrix at moment $k+1$ are given by

$$\hat{\mathbf{x}}_{k+1|k} = \mathbf{A_d}\hat{\mathbf{x}}_{k|k} + \mathbf{B}\mathbf{h}(\mathbf{z}^k) + \mathbf{B}\mathbf{e}[k], \tag{16}$$

$$\mathbf{P}_{k+1|k} = \mathbf{A_d}\mathbf{P}_{k|k}{\mathbf{A_d}}^T + \mathbf{\Sigma_w}. \tag{17}$$

Define $\boldsymbol{\upsilon}[k+1] \in \mathcal{R}^p$ as the innovation vector at moment $k+1$ as

$$\boldsymbol{\upsilon}[k+1] = \mathbf{z}[k+1] - \mathbf{C}\hat{\mathbf{x}}_{k+1|k}, \tag{18}$$

and the corresponding innovation covariance matrix is

$$\mathbf{R}_{k+1} = \mathbf{C}\mathbf{P}_{k+1|k}\mathbf{C}^T + \mathbf{\Sigma_n}. \tag{19}$$

The optimal Kalman gain matrix at moment $k+1$ is

$$\mathbf{K}_{k+1} = \mathbf{P}_{k+1|k}\mathbf{C}^T\mathbf{R}_{k+1}^{-1}. \tag{20}$$

Then the posterior state estimation and the corresponding covariance matrix at moment $k+1$ are updated by

$$\hat{\mathbf{x}}_{k+1|k+1} = \hat{\mathbf{x}}_{k+1|k} + \mathbf{K}_{k+1}\boldsymbol{\upsilon}[k+1], \tag{21}$$

$$\mathbf{P}_{k+1|k+1} = (\mathbf{I}_p - \mathbf{K}_{k+1}\mathbf{C})\mathbf{P}_{k+1|k}. \tag{22}$$

Substituting (16) and (18) into (21) yields

$$\begin{aligned}\hat{\mathbf{x}}_{k+1|k+1} = {} & \mathbf{A_d}\hat{\mathbf{x}}_{k|k} + \mathbf{B}\mathbf{h}(\mathbf{z}^k) \\ & + \mathbf{B}\mathbf{e}[k] + \mathbf{K}_{k+1}\boldsymbol{\upsilon}[k+1].\end{aligned} \tag{23}$$

B. Statistical Test with Dynamic Watermarking

Based on the state estimation results in (23), define the additive distortion power of the attacker at moment $k+1$ as

$$\begin{aligned}\mathbf{d}[k+1] = {} & \hat{\mathbf{x}}_{k+1|k+1} - \mathbf{A_d}\hat{\mathbf{x}}_{k|k} - \mathbf{B}\mathbf{h}(\mathbf{z}^k) \\ & -\mathbf{B}\mathbf{e}[k] - \mathbf{K}_{k+1}\boldsymbol{\upsilon}[k+1].\end{aligned} \tag{24}$$

If there is no attack, then $\mathbf{d}[k+1] = \mathbf{0}$, and we have the following distributions,

$$\begin{aligned}&\hat{\mathbf{x}}_{k+1|k+1} - \mathbf{A_d}\hat{\mathbf{x}}_{k|k} - \mathbf{B}\mathbf{h}(\mathbf{z}^k) \sim \\ &\quad \mathcal{N}_n(\mathbf{0}, \mathbf{B}\mathbf{\Sigma_e}\mathbf{B}^T + \mathbf{K}_{k+1}\mathbf{R}_{k+1}\mathbf{K}_{k+1}^T),\end{aligned} \tag{25}$$

$$\begin{aligned}&\hat{\mathbf{x}}_{k+1|k+1} - \mathbf{A_d}\hat{\mathbf{x}}_{k|k} - \mathbf{B}\mathbf{h}(\mathbf{z}^k) - \mathbf{B}\mathbf{e}[k] \sim \\ &\quad \mathcal{N}_n(\mathbf{0}, \mathbf{K}_{k+1}\mathbf{R}_{k+1}\mathbf{K}_{k+1}^T).\end{aligned} \tag{26}$$

Define a test statistic $\mathbf{g}[k+1]$ at moment $k+1$ as the sum of the attack power vector and a scaled innovation vector as

$$\begin{aligned}\mathbf{g}[k+1] &= \mathbf{d}[k+1] + \mathbf{K}_{k+1}\boldsymbol{\upsilon}[k+1] \\ &= \hat{\mathbf{x}}_{k+1|k+1} - \mathbf{A_d}\hat{\mathbf{x}}_{k|k} - \mathbf{B}\mathbf{h}(\mathbf{z}^k) - \mathbf{B}\mathbf{e}[k].\end{aligned} \tag{27}$$

The elements in $\mathbf{g}[k+1]$ might be mutually correlated because of the selected state of the system, which makes $\mathbf{\Phi}_{k+1} = \mathbf{K}_{k+1}\mathbf{R}_{k+1}\mathbf{K}_{k+1}^T$ singular.

To solve this problem, denote $\mathbf{\Phi} = \lim_{k\to\infty}\mathbf{\Phi}_{k+1}$ as the asymptotic estimate of $\mathbf{\Phi}_{k+1}$. Assume the rank of $\mathbf{\Phi}$ is $q \le p$ with nonzero eigenvalues $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_q]^T$, and the matrix $\bar{\mathbf{U}} \in \mathcal{C}^{p\times q}$ contains the corresponding eigenvectors on its column. We can perform dimension reduction on $\mathbf{g}[k+1]$ as

$$\bar{\mathbf{g}}[k+1] = \bar{\mathbf{U}}^H\mathbf{g}[k+1]. \tag{28}$$

Then we have

$$\lim_{k\to\infty}\mathbb{E}\left[\bar{\mathbf{g}}[k+1]\bar{\mathbf{g}}[k+1]^H\right] = \mathbf{D}, \tag{29}$$

where $\mathbf{D} = \mathrm{Diag}(\boldsymbol{\lambda}) \in \mathcal{C}^{q\times q}$ is a diagonal matrix with the q nonzero eigenvalues of $\mathbf{\Phi}$ on its main diagonal.

Based on the test statistic, the statistical tests that are used for dynamic watermarking are [21]

1) Test 1:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{e}[k]\bar{\mathbf{g}}[k+1]^T = \mathbf{0}_{m\times q}. \tag{30}$$

2) Test 2:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\bar{\mathbf{g}}[k+1]\bar{\mathbf{g}}[k+1]^T = \mathbf{D}. \tag{31}$$

Tests 1 and 2 correspond to the distributions given in (25) and (26), respectively. Test 1 is used to test the independence between the watermark signal, $\mathbf{e}$, and the test statistic, $\bar{\mathbf{g}}$. Test 2 is used to ensure the measurements conform to the state estimation obtained from the Kalman filter. Without the knowledge of the dynamic watermark, a falsified measurement cannot pass both Tests 1 and 2. Thus both tests are indispensable for the active detection process.

Following the similar procedure as in [21], it can be proved that passing both tests is sufficient to achieve an asymptotically zero attacking power as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}||\mathbf{d}[k+1]||^2 = 0, \tag{32}$$

which means there is no attack.

This two-test detection procedure can be simplified into one test by combining the two statistics $\mathbf{e}[k]$ and $\bar{\mathbf{g}}[k+1]$ into one vector [23]. Define

$$\mathbf{r}[k+1] = [\bar{\mathbf{g}}[k+1]^T,\ \mathbf{e}[k]^T]^T \in \mathcal{R}^{q+m}. \tag{33}$$

Then the the two tests described in (30) and (31) can be combined into one equivalent test as,

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{r}[k+1]\mathbf{r}[k+1]^T, \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\begin{bmatrix}\bar{\mathbf{g}}[k+1]\bar{\mathbf{g}}[k+1]^T & \bar{\mathbf{g}}[k+1]\mathbf{e}[k]^T \\ \mathbf{e}[k]\bar{\mathbf{g}}[k+1]^T & \mathbf{e}[k]\mathbf{e}[k]^T\end{bmatrix}, \\ &= \begin{bmatrix}\mathbf{D} & \mathbf{0}_{q\times m} \\ \mathbf{0}_{m\times q} & \mathbf{\Sigma_e}\end{bmatrix} := \mathbf{\Sigma_0} \in \mathcal{R}^{(q+m)\times(q+m)}.\end{aligned} \tag{34}$$

IV. Low Latency Detection with Dynamic Watermarking

Building upon the test statistics crafted for dynamic watermarking, we propose to develop a low-latency attack detection methods under the framework of dynamic watermarking.

A. Low Latency Attack Detection

The objective of low latency attack detection is to minimize the average detection delay (ADD) subject to an upper bound of the probability of false alarm (PFA). Denote the attack time identified by the detector as $\hat{\tau}$. Then the detection problem can be formulated as,

$$\begin{aligned} \min \quad & \text{ADD} = \mathbb{E}[\hat{\tau} - \tau | \hat{\tau} > \tau], \\ \text{s.t.} \quad & \text{PFA} = \text{P}(\hat{\tau} < \tau) \leq \alpha. \end{aligned} \tag{P1}$$

There is a fundamental tradeoff between ADD and PFA [16]. The ADD can be reduced at the cost of a higher PFA, and vice versa. The problem formulation in (P1) aims to minimize the detection delay, subject to an upper bound on the PFA to achieve guaranteed detection accuracy.

Based on the analysis in the previous section, the distributions of the dynamic watermark test statistic, $\mathbf{r}[k+1]$, under the null and alternative hypothesis can be represented as

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{r}[k+1] \sim \mathcal{N}_{q+m}(\mathbf{0}, \mathbf{\Sigma_0}), \\ \mathcal{H}_1 &: \mathbf{r}[k+1] \sim \mathcal{N}_{q+m}(\mu, \mathbf{\Sigma}), \end{aligned} \tag{35}$$

where μ and $\mathbf{\Sigma}$ are the post-attack mean and covariance matrix, respectively. The values of μ and $\mathbf{\Sigma}$ for different attacks are analyzed in the next subsection.

Define a new variable $\Gamma[k]$ as

$$\Gamma[k] = \mathbf{r}[k]^T {\mathbf{\Sigma_0}}^{-1} \mathbf{r}[k]. \tag{36}$$

Under the null hypothesis, $\Gamma[k]$ follows a χ^2-distribution with $q+m$ degrees of freedom with mean and variance given as follows,

$$\begin{aligned} \mathbb{E}[\Gamma[k]] &= q+m \\ \text{Var}[\Gamma[k]] &= 2(q+m) \end{aligned} \tag{37}$$

Based on the distribution of $\Gamma[k]$, we can define the test statistics used for CUSUM as [8]

$$U[k+1] = \max(0, U[k] + \frac{\Gamma[k+1] - (q+m)}{\sqrt{2(q+m)}}), \tag{38}$$

$$T[k] = \frac{U[k]}{k} \tag{39}$$

with $U[1] = 0$. The test sequence $T[k]$ accumulates the normalized variable, $\frac{\Gamma[k]-(q+m)}{\sqrt{2(q+m)}}$, over time. Under the null hypothesis, the test sequence $T[k]$ is always close to 0. Under the event of cyberattacks, the value of $T[k]$ will increase over time. Thus the CUSUM detector can be defined as a threshold test as

$$\hat{\tau} = \inf\{k \geq 1 | T[k] \geq \alpha\}, \tag{40}$$

where the threshold α is chosen to meet the PFA upper bound constraint. The Markov chain approach in [36] can be used for calculating the PFA and select the threshold.

B. Post-Attack Distributions and KL Divergence

We propose to measure the stealthiness of different attacks by using the Kullback-Leibler (KL) divergence between the pre- and post-attack distributions of $\mathbf{r}[k+1]$. The KL divergence is a measure about how one probability distribution is different from a second one. A smaller KL divergence between two probability distributions means that the two distributions are similar to each other, thus it will be harder to distinguish between the two. In terms of low latency attack detection, it has been shown that the detection delay is inversely proportional to the KL divergence between the two distributions before and after the attack [16, Theorem 3.1]. Thus a smaller KL divergence means a longer detection delay, which corresponds to a stealthier attack.

The KL divergence of the pre- and post-attack distributions can be calculated as

$$D(\mathcal{H}_1 || \mathcal{H}_0) = \frac{1}{2}[\mu^{\mathbf{T}} \mathbf{\Sigma_0^{-1}} \mu + \text{Tr}(\mathbf{\Sigma_0^{-1}} \mathbf{\Sigma}) + \log \frac{|\mathbf{\Sigma_0}|}{|\mathbf{\Sigma}|} - m - q], \tag{41}$$

with μ and $\mathbf{\Sigma}$ being the post-attack mean and covariance matrices for the various attacks.

The calculations of the KL divergence requries the knowledge of the pre- and post-attack distributions of $\mathbf{r}[k]$, which are analyzed as follows.

Under normal operation conditions without any attack, the limit distribution of $\mathbf{r}[k+1]$ is given based on the Law of large numbers (LLN) as

$$\lim_{k \to \infty} \mathbf{r}[k+1] \sim \mathcal{N}_{q+m}(\mathbf{0}, \mathbf{\Sigma_0}). \tag{42}$$

Denote $\mathbf{K} = \lim_{k\to\infty} \mathbf{K}_{k+1}$ and $\mathbf{P} = \lim_{k\to\infty} \mathbf{P}_{k+1}$ as the asymptotic covariance matrix and Kalman gain matrix, respectively. The post-attack distribution of $\mathbf{r}[k+1]$ depends on the various attack models as analyzed in the following.

1) FDI attack: Substituting (9) and (18) into (24) and (27), we have the post-attack distribution of $\mathbf{r}[k+1]$ under the FDI attack as

$$\lim_{k \to \infty} \mathbf{r}[k+1] \sim \mathcal{N}_{q+m}(\mu, \mathbf{\Sigma}) \tag{43}$$

with

$$\mu = \begin{bmatrix} \bar{\mathbf{U}}^H \mathbf{K} \mathbf{a}[k+1] \\ \mathbf{0}_m \end{bmatrix} \tag{44a}$$

$$\mathbf{\Sigma} = \mathbf{\Sigma_0} \tag{44b}$$

under deterministic FDI. Under the noisy FDI attack, we have

$$\mu = \mathbf{0}_{q+m} \tag{45a}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{D} + \bar{\mathbf{U}}^H \mathbf{K} \mathbf{\Sigma_a} \mathbf{K}^T \bar{\mathbf{U}} & \mathbf{0}_{q \times m} \\ \mathbf{0}_{m \times q} & \mathbf{\Sigma_e} \end{bmatrix} \tag{45b}$$

under noise FDI.

2) Replay attack: Substitute (10) to (24) and (27). Define the control matrix $\mathbf{L}$ to be the linear approximation of the control policy $\mathbf{h}(\cdot)$, such that

$$\mathbf{u}[k] = \mathbf{L} \hat{\mathbf{x}}_{k|k} + \mathbf{e}[k]. \tag{46}$$

Then the post-attack distribution of $\mathbf{r}[k+1]$ under the replay attack can be estimated as,

$$\lim_{k\to\infty} \mathbf{r}[k+1] \sim \mathcal{N}_{q+m}(\mathbf{0}_{q+m}, \mathbf{\Sigma}), \tag{47}$$

with

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{D} + 2\bar{\mathbf{U}}^H\mathbf{KCXC^T}\mathbf{K}^T\bar{\mathbf{U}} & -\bar{\mathbf{U}}^\mathbf{H}\mathbf{KCB}\mathbf{\Sigma_e} \\ -\mathbf{\Sigma_e}\mathbf{B}^T\mathbf{C}^T\mathbf{K}^T\bar{\mathbf{U}} & \mathbf{\Sigma_e}, \end{bmatrix} \tag{48}$$

where $\mathbf{X}$ is the solution of the following Lyapunov equation

$$\mathbf{A_e X A_e^T} - \mathbf{X} + \mathbf{B}\mathbf{\Sigma_e}\mathbf{B^T} = \mathbf{0}, \tag{49}$$

and $\mathbf{A_e}$ is the estimated transition matrix:

$$\mathbf{A_e} = (\mathbf{A_d} + \mathbf{BL})(\mathbf{I}_p - \mathbf{KC}). \tag{50}$$

3) Destabilization attack: Substituting (8) into (23) leads to the post-attack distribution as

$$\mathbf{r}[k+1] \sim \mathcal{N}_{q+m}(\mu, \mathbf{\Sigma}), \tag{51}$$

with

$$\mu = \begin{bmatrix} \bar{\mathbf{U}}^H\mathbf{KCBA_p}\hat{\mathbf{x}}_{k|k} \\ \mathbf{0_m}, \end{bmatrix} \tag{52a}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{D} + \bar{\mathbf{U}}^H\mathbf{KCP_a}\mathbf{C}^T\mathbf{K}^T\bar{\mathbf{U}} & \mathbf{0}_{q\times m} \\ \mathbf{0}_{m\times q} & \mathbf{\Sigma_e}, \end{bmatrix} \tag{52b}$$

where

$$\mathbf{P_a} = \mathbf{BA_p P A_p}^T\mathbf{B}^T + \mathbf{A_d P A_p}^T\mathbf{B}^T + \mathbf{BA_p P A_d}^T. \tag{53}$$

V. Simulation Results

Simulation results are presented in this section to verify the performance of the proposed low latency attack detection method. The PV farm model shown in Fig. 1 is implemented by using Matlab Simulink. All attack simulations are performed by using the Simulink model. In the simulation, the DC link voltage V_{DC} is set at 800 V, and the magnitude of the output AC phase voltage $|V_{\text{G}}|$ is set as $400 \times \sqrt{\frac{2}{3}} = 326.60$ V, operating at a frequency of 60 Hz. The reference DQ frame currents are defined as $I_{\text{D}}^{\text{Ref}} = -150$ A and $I_{\text{Q}}^{\text{Ref}} = 0$ A. The proportional parameters K_{p1} and K_{p2} are both set to 10, and the integral parameters K_{i1} and K_{i2} are adjusted to 20.

The simulation time interval is set to $\Delta t = 10^{-6}$ s, which corresponds to a sampling rate of 1 MHz. The continuous state-space model is discretized using a 2 kHz sampling rate, corresponding to a time interval of 5×10^{-4} s between measurements. The covariance matrices for process and measurement noises are respectively defined as $\mathbf{\Sigma_w} = 10^{-6}\mathbf{I}_4$ and $\mathbf{\Sigma_n} = 5\times10^{-7}\mathbf{I}_2$. The covariance matrix for the dynamic watermark is $\mathbf{\Sigma_e} = 10^{-6}\mathbf{I}_2$.

Equilibrium is achieved within 2 seconds during simulations. Once the system reaches the equilibrium, data are collected during the next minute for parameter estimation. The state $\mathbf{x}$, the input $\mathbf{u}$, and the output $\mathbf{y}$ in the one minute period are recorded, and are then used to estimate the corresponding matrices $\mathbf{A_d}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}, \mathbf{K}, \mathbf{P}$.

Cyberattacks and low latency attack detection are performed after parameter estimations. The attacks are launched at 4.5 s after the parameter estimation. State estimations are performed by using the control inputs and the measurements, and the results are then used to calculate the CUSUM test statistic. The ADD and PFA for the detector are computed using results gathered from 1,000 Monte Carlo simulation trials.

A. Deterministic FDI Attack

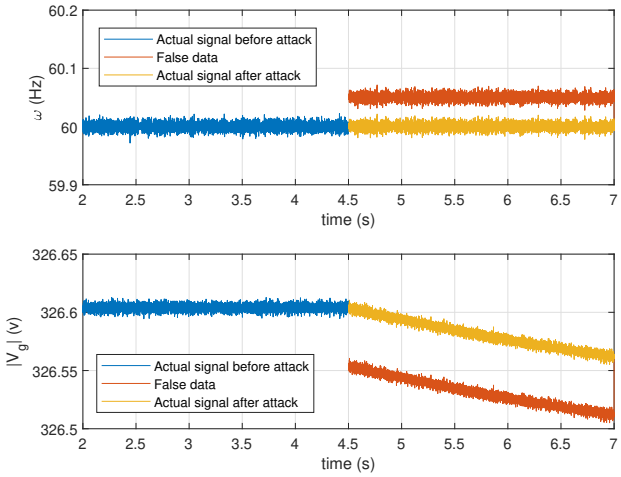

Fig. 2: The voltage frequency (top) and magnitude (bottom) measurement under deterministic FDI attack on the PV system at $4.5s$

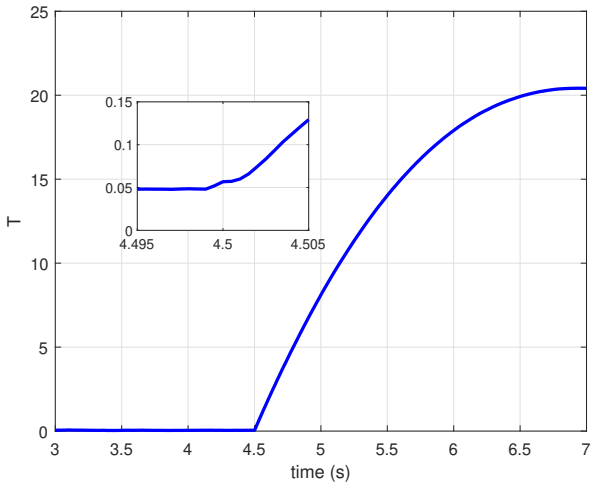

Fig. 3: The detector statistic under deterministic FDI attack on the PV system at $4.5s$

The FDI attack with a deterministic attack vector is simulated by injecting the vector $\mathbf{a}[t] = [0.05, -0.05]^T$ to the measurement vector at 4.5 s. Fig. 2 shows the actual measurements and those under attack. Even though the attack does not

lead to an apparent frequency deviation, the voltage magnitude drops due to the attack.

The CUSUM statistic under the deterministic FDI attack with a zoom-in around 4.5s is presented in Fig. 3. The CUSUM statistic is around 0 prior to the attack, and its value increases dramatically after the attack. Thus the attack can be easily detected with minimum delay with the proposed low latency attack detection algorithm.

B. Noisy FDI Attack

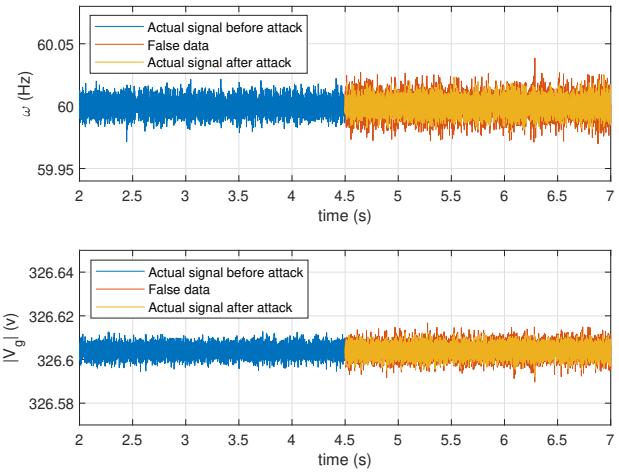

Fig. 4: The voltage frequency (top) and magnitude (bottom) measurement under noise FDI attack on the PV system at $4.5s$

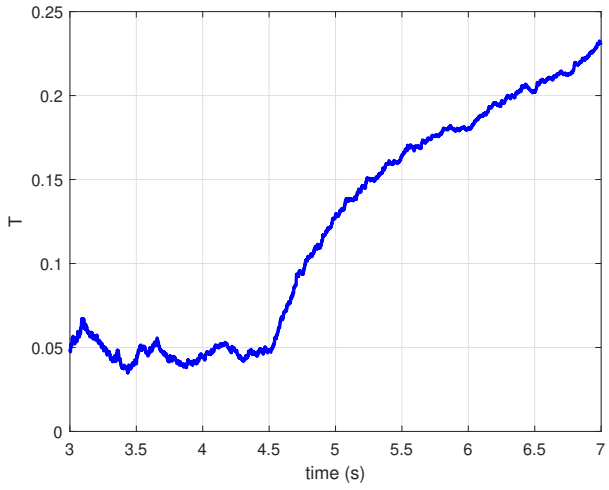

Fig. 5: The detector statistic under noise FDI attack on the PV system at $4.5s$

The noisy FDI attack vector is generated from a random vector $\mathbf{a}[t] \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma_a})$. We set the noise covariance to a level that is multiple times of the system and measurement noise, i.e., the vector on the main diagonal of $\mathbf{\Sigma_a}$ is set to $[3\times10^{-5}, 3\times10^{-6}]$. Fig. 4 shows the actual measurements and those with noise injections. The injection only causes trivial fluctuation in both the frequency and voltage magnitude, and the actual measurements still fall in a normal range because of the control system.

The CUSUM statistic under the noisy FDI attack is presented in Fig. 5. Since the variance of the injected noise is very low, such an attack is hard to detect. However, it still causes a significant increase in the slope of the CUSUM statistic. Thus the noisy FDI attack can be easily detected with the proposed detection algorithm with low detection latency.

C. Replay attack

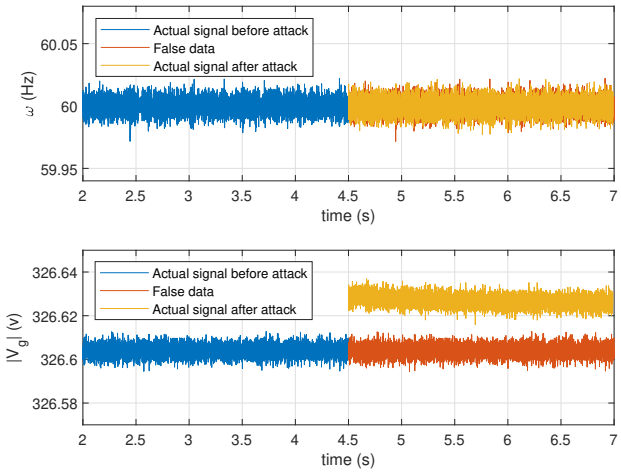

Fig. 6: The voltage frequency (top) and magnitude (bottom) measurement under replay attack on the PV system at $4.5s$

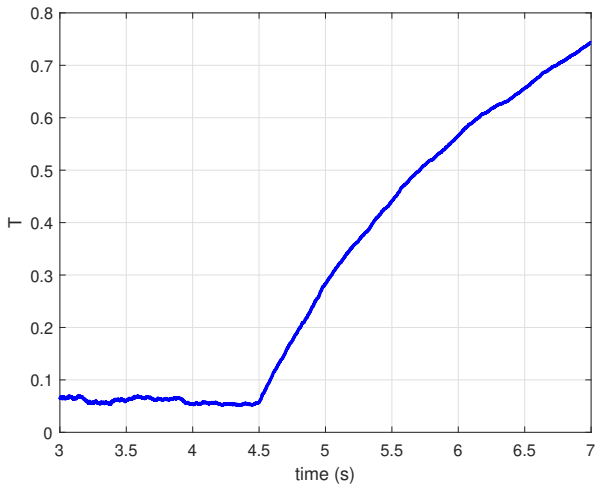

Fig. 7: The detector statistic under replay attack on the PV system at $4.5s$

The replay attack is simulated by replacing the actual measurements from 4.5 s by historical measurements starting at 2.5 s (a delay of 2 seconds). Fig. 6 shows the measurements between 2s to 7s, where there is a 2 second delay between the attacked measurement and the actual measurement. The replay

attack does not deviate the measurements from their normal range. However, the system will be out of normal control and cannot respond to load changes of the grid or faults in the PV farm, which can cause voltage fluctuations, reverse power flow, and real power curtailments.

The CUSUM statistic under the replay attack is presented in Fig. 7. The statistic $T[k]$ increases much slower than other attacks, i.e., the attack is much stealthier compared to others. However, there is still an apparent increase in the slope of $T[k]$. Thus the replay attack can be easily detected with the proposed algorithm even if it does not cause significant deviations of the system states.

D. Destabilization attack

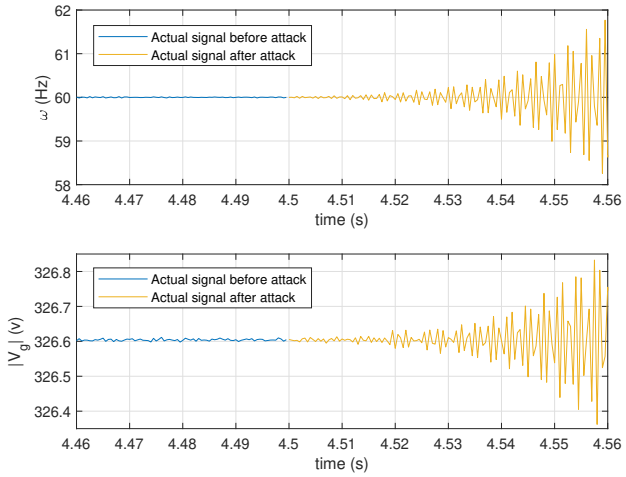

Fig. 8: The voltage frequency (top) and magnitude (bottom) measurement under destabilization attack on the PV system at $4.5s$

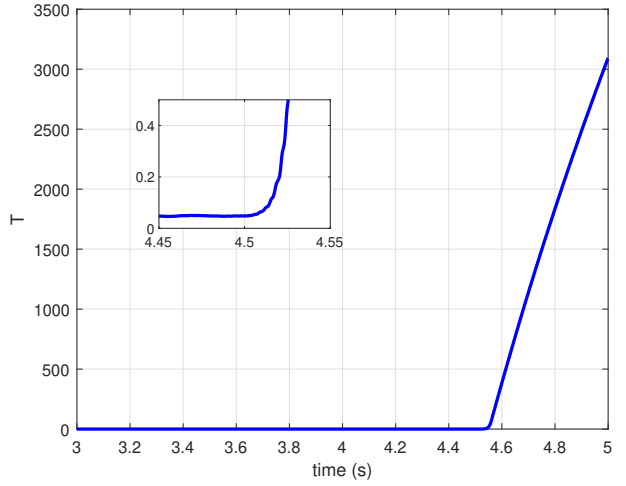

Fig. 9: The detector statistic under destabilization attack on the PV system at $4.5s$

The destabilization attack is launched by replacing the control inputs $\mathbf{u}[t]$ with $\mathbf{u}[t] + \mathbf{A_p x}[t]$ starting at 4.5 s, where $\mathbf{A_p}$ is a diagonal matrix with the main diagonal being $[1.5, 1.5, 0, 0]$, such that $||\mathbf{A_d} + \mathbf{BA_p}|| \geq 1$. Fig. 8 shows the measurements from 4.46 s to 4.56 s. The attack rapidly causes instability in measurements which gradually exceeds its normal range.

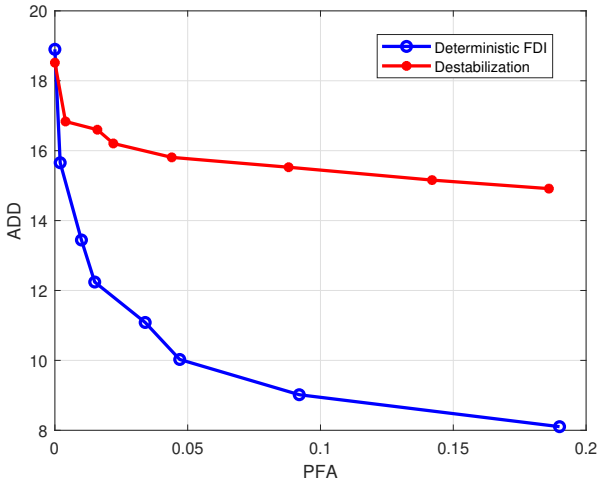

Fig. 10: The PFA-ADD curve under the deterministic FDI attack and destabilization attack on the PV system

The CUSUM statistic under destabilization attack is presented in Fig. 9. The statistic $T[k]$ increases much faster than other attacks, such that it is easier to detect such an attack.

E. Detector performance

More powerful attacks can make the system rapidly drift away from its normal state and cause damage in a short period of time. However, they are usually easier to detect. The adversaries have more incentives to balance the stealthiness and power of the attack such that they can cause damages before being detected.

The stealthiness of the attacks can be measured by using the KL-divergence between the distributions of the CUSUM test vector $\mathbf{r}$ before and after the attack. The KL-divergence of various attacks at 4.5s is calculated by using the results in Section IV-B, and the results are shown in Table I. The deterministic FDI and destabilization attacks have similar levels of KL divergence, and both are two or three orders of magnitude higher than that of the relay and noisy FDI attacks. Thus the deterministic FDI and destabilization attacks are relatively easier to detect. Among the 4 attacks, the noisy FDI attack has the best stealthiness with the lowest KL divergence.

The performance of the proposed low latency CUSUM detector is evaluated by using the ADD-PFA tradeoff curves shown in Figs. 10 and 11. Each point on the ADD-PFA tradeoff curve is obtained through 1,000 Monte Carlo trials for a given detection threshold. Under the same PFA, e.g. PFA = 0.02, the ADD of the deterministic FDI, destabilization, replay, and noisy attacks are 12.1, 16.5, 109, and 128 ms, respectively. This is consistent with the KL divergence results, that is, attacks with lower KL divergence are harder to detect, thus they have larger ADD under the same PFA.

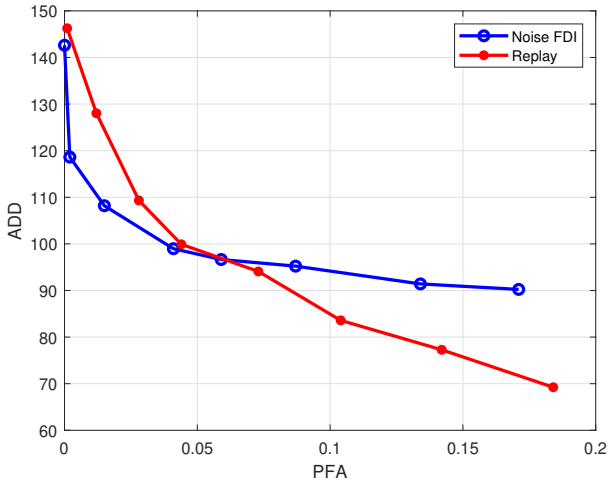

Fig. 11: The PFA-ADD curve under noise FDI attack and replay attack on the PV system

Attack type	KL-divergence
FDI (deterministic)	0.2979
FDI (noise)	0.0053
Replay	0.0547
Destabilization	0.4567

TABLE I: KL-divergence between distribution of $\mathbf{r}$ before and after attack.

VI. Conclusion

This paper has proposed an active low latency attack detection algorithm for grid-connected PV systems. We have developed a generalized CUSUM detector with dynamic watermarking by constructing and analyzing the physical model of a grid-connected PV system. The detection algorithm was developed to minimize detection delay while ensuring detection accuracy. In addition, we have proposed to use the KL divergence to measure the stealthiness of different cyberattacks. The algorithm was tested on a 400 V grid-connected PV system with various cyberattacks. Simulation results demonstrated that the proposed algorithm can achieve a detection delay of 50 ms with PFA below 5%.

Appendix A
Proof of Equation (44) and (45)

Under the FDI attack, the measurements are replaced by:

$$\mathbf{z}[k+1] = \mathbf{y}[k+1] + \mathbf{a}[k+1] \tag{54}$$

The posterior state estimation is:

$$\begin{aligned}&\hat{\mathbf{x}}_{k+1|k+1}\\ &= \hat{\mathbf{x}}_{k+1|k} + \mathbf{K}_{k+1}(\mathbf{z}[k+1] - \mathbf{C}\hat{\mathbf{x}}_{k+1|k})\\ &= \hat{\mathbf{x}}_{k+1|k} + \mathbf{K}_{k+1}(\mathbf{y}[k+1] - C\hat{\mathbf{x}}_{k+1|k} + \mathbf{a}[k+1])\end{aligned} \tag{55}$$

The test statistic $\mathbf{g}[k+1]$ is:

$$\begin{aligned}\mathbf{g}[k+1] &= \hat{\mathbf{x}}_{k+1|k+1} - \hat{\mathbf{x}}_{k+1|k}\\ &= \mathbf{K}_{k+1}(\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1])\end{aligned} \tag{56}$$

Then the whitened statistic $\bar{\mathbf{g}}[k+1]$ is

$$\bar{\mathbf{g}}[k+1] = \bar{\mathbf{U}}^H\mathbf{g}[k+1] = \bar{\mathbf{U}}^H\mathbf{K}_{k+1}(\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1]) \tag{57}$$

For deterministic $\mathbf{a}[k+1]$, the mean of $\bar{\mathbf{g}}[k+1]$:

$$\mathbb{E}[\bar{\mathbf{g}}[k+1]] = \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1]] = \bar{\mathbf{U}}^H\mathbf{K}\mathbf{a}[k+1] \tag{58}$$

The covariance of $\bar{\mathbf{g}}[k+1]$ is $\mathbf{D}$ since the mean is deterministic:

$$\mathrm{Cov}[\bar{\mathbf{g}}[k+1]] = \mathbf{D} \tag{59}$$

The covariance of $\bar{\mathbf{g}}[k+1]$ and $\mathbf{e}[k]$ is:

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{g}}[k+1]\mathbf{e}[k]^T] &= \mathbb{E}[\bar{\mathbf{U}}^H\mathbf{K}_{k+1}(\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1])\mathbf{e}[k]^T]\\ &= \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[(\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1])\mathbf{e}[k]^T]\\ &= \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[(\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1])]\mathbb{E}[\mathbf{e}[k]^T]\\ &= \mathbf{0}_{q\times m}\end{aligned} \tag{60}$$

Then the mean of $\mathbf{r}[k+1]$ is:

$$\mu = \begin{bmatrix}\mathbb{E}[\bar{\mathbf{g}}[k+1]]\\ \mathbb{E}[\mathbf{e}[k]]\end{bmatrix} = \begin{bmatrix}\bar{\mathbf{U}}^H\mathbf{K}\mathbf{a}[k+1]\\ \mathbf{0}_m\end{bmatrix} \tag{61}$$

and the covariance of $\mathbf{r}[k+1]$ is:

$$\Sigma = \begin{bmatrix}\mathbf{D} & \mathbf{0}_{q\times m}\\ \mathbf{0}_{m\times q} & \boldsymbol{\Sigma}_\mathbf{e}\end{bmatrix} = \boldsymbol{\Sigma}_\mathbf{0} \tag{62}$$

which completes the proof of (44).

For $\mathbf{a}[k+1] \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{a})$, the mean of $\bar{\mathbf{g}}[k+1]$ is:

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{g}}[k+1]] &= \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1]]\\ &= \bar{\mathbf{U}}^H\mathbf{K}(\mathbb{E}[\boldsymbol{\upsilon}[k+1] + \mathbb{E}[\mathbf{a}[k+1]]) = \mathbf{0}_q\end{aligned} \tag{63}$$

The covariance of $\bar{\mathbf{g}}[k+1]$ is:

$$\begin{aligned}&\mathrm{Cov}[\bar{\mathbf{g}}[k+1]] = \mathbb{E}[\bar{\mathbf{g}}[k+1]\bar{\mathbf{g}}[k+1]^T]\\ &= \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[(\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1])(\boldsymbol{\upsilon}[k+1] + \mathbf{a}[k+1])^T]\mathbf{K}^T\bar{\mathbf{U}}\\ &= \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[\boldsymbol{\upsilon}[k+1]\boldsymbol{\upsilon}[k+1]^T + \mathbf{a}[k+1]\mathbf{a}[k+1]^T]\mathbf{K}^T\bar{\mathbf{U}}\\ &= \mathbf{D} + \bar{\mathbf{U}}^H\mathbf{K}\boldsymbol{\Sigma}_\mathbf{a}\mathbf{K}^T\bar{\mathbf{U}}\end{aligned} \tag{64}$$

using the fact that innovation $\boldsymbol{\upsilon}[k+1]$ and $\mathbf{a}[k+1]$ are independent. The covariance of $\bar{\mathbf{g}}[k+1]$ and $\mathbf{e}[k]$ has the same form as deterministic case, then the mean and covariance of $\mathbf{r}[k+1]$ is:

$$\mu = \mathbf{0}_{q+m} \tag{65a}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix}\mathbf{D} + \bar{\mathbf{U}}^H\mathbf{K}\boldsymbol{\Sigma}_\mathbf{a}\mathbf{K}^T\bar{\mathbf{U}} & \mathbf{0}_{q\times m}\\ \mathbf{0}_{m\times q} & \boldsymbol{\Sigma}_\mathbf{e}\end{bmatrix} \tag{65b}$$

which completes the proof of (45).

APPENDIX B
PROOF OF EQUATION (47) AND (48)

Under the replay attack, the measurements are replaced by:

$$\mathbf{z}[k+1] = \mathbf{y}[k+1-l] \tag{66}$$

The innovation after attack is:

$$\begin{aligned}\boldsymbol{\upsilon}[k+1] &= \mathbf{z}[k+1] - \mathbf{C}\hat{\mathbf{x}}_{k+1|k} \\ &= \mathbf{y}[k+1-l] - \mathbf{C}\hat{\mathbf{x}}_{k+1|k} \\ &= \mathbf{C}(\mathbf{x}[k+1-l] - \hat{\mathbf{x}}_{k+1|k}) + \mathbf{n}[k+1-l] \\ &= \mathbf{C}(\mathbf{x}[k+1-l] - \mathbf{A_d}\hat{\mathbf{x}}_{k|k} - \mathbf{Bh}(\mathbf{z}^k)) \\ &\quad + \mathbf{n}[k+1-l] - \mathbf{CBe}[k]\end{aligned} \tag{67}$$

Suppose we have a virtual system that satisfied the following system equations:

$$\mathbf{x}^{'}[t+1] = \mathbf{A_d}\mathbf{x}^{'}[t] + \mathbf{Bh}(\mathbf{z}^{'t}) + \mathbf{Be}^{'}[t] + \mathbf{w}^{'}[t+1], \tag{68a}$$

$$\mathbf{y}^{'}[t+1] = \mathbf{C}\mathbf{x}^{'}[t+1] + \mathbf{n}^{'}[t+1]. \tag{68b}$$

and the Kalman filter update at $k+1$:

$$\hat{\mathbf{x}}^{'}_{k+1|k} = \mathbf{A_d}\hat{\mathbf{x}}^{'}_{k|k} + \mathbf{Bh}(\mathbf{z}^{'k}) + \mathbf{Be}^{'}[k] \tag{69}$$

$$\hat{\mathbf{x}}^{'}_{k+1|k+1} = \hat{\mathbf{x}}^{'}_{k+1|k} + \mathbf{K}_{k+1}\boldsymbol{\upsilon}^{'}[k+1] \tag{70}$$

and also satisfies the linear approximation of control policy:

$$\mathbf{h}(\mathbf{z}^{'k}) = \mathbf{L}\hat{\mathbf{x}}^{'}_{k|k} \tag{71}$$

with the initial state $\mathbf{x}^{'}[0]$ and initial prior state estimation $\hat{\mathbf{x}}^{'}_{1|0}$. In addition, the virtual system is a delayed version of the real system without any attack, which satisfies:

$$\mathbf{x}^{'}[k+1] = \mathbf{x}[k+1-l] \tag{72}$$

$$\hat{\mathbf{x}}^{'}_{k|k} = \hat{\mathbf{x}}_{k-l|k-l} \tag{73}$$

when $k \geq l$. Then the replay attack can be regarded as replacing $\mathbf{y}[k]$ with $\mathbf{y}^{'}[k]$ starting from τ.

Define the estimated control and transition matrix:

$$\mathbf{u}[k] := \mathbf{L}\hat{\mathbf{x}}_{k|k} + \mathbf{e}[k] \tag{74}$$

$$\mathbf{A_e} := (\mathbf{A_d} + \mathbf{BL})(\mathbf{I}_p - \mathbf{KC}) \tag{75}$$

Assume $\mathbf{A_e}$ is stable, otherwise the measurements will soon be unbounded and the attack can be detected as a destabilization attack.

The Kalman filter estimation after the system is attacked and becomes stable can be rewritten as:

$$\begin{aligned}&\hat{\mathbf{x}}_{k+1|k} = \mathbf{A_d}\hat{\mathbf{x}}_{k|k} + \mathbf{Bu}[k] \\ &= (\mathbf{A_d} + \mathbf{BL})\hat{\mathbf{x}}_{k|k} + \mathbf{Be}[k] \\ &= (\mathbf{A_d} + \mathbf{BL})(\hat{\mathbf{x}}_{k|k-1} + \mathbf{K}(\mathbf{y}^{'}[k] - \mathbf{C}\hat{\mathbf{x}}_{k|k-1}) + \mathbf{Be}[k] \\ &= \mathbf{A_e}\hat{\mathbf{x}}_{k|k-1} + (\mathbf{A_d} + \mathbf{BL})\mathbf{K}\mathbf{y}^{'}[k] + \mathbf{Be}[k]\end{aligned} \tag{76}$$

This update also holds true for the virtual system that:

$$\hat{\mathbf{x}}^{'}_{k+1|k} = \mathbf{A_e}\hat{\mathbf{x}}^{'}_{k|k-1} + (\mathbf{A_d} + \mathbf{BL})\mathbf{K}\mathbf{y}^{'}[k] + \mathbf{Be}^{'}[k] \tag{77}$$

Therefore, we consider the difference between the prior estimation of the two systems at $k+1$:

$$\begin{aligned}&\hat{\mathbf{x}}^{'}_{k+1|k} - \hat{\mathbf{x}}_{k+1|k} \\ &= \mathbf{A_e}(\hat{\mathbf{x}}^{'}_{k|k-1} - \hat{\mathbf{x}}_{k|k-1}) + \mathbf{B}(\mathbf{e}^{'}[k] - \mathbf{e}[k]) \\ &= {\mathbf{A_e}}^2(\hat{\mathbf{x}}^{'}_{k-1|k-2} - \hat{\mathbf{x}}_{k-1|k-2}) \\ &\quad + \mathbf{A_e}\mathbf{B}(\mathbf{e}^{'}[k-1] - \mathbf{e}[k-1]) + \mathbf{B}(\mathbf{e}^{'}[k] - \mathbf{e}[k])) \\ &= \cdots \\ &= {\mathbf{A_e}}^k(\hat{\mathbf{x}}^{'}_{1|0} - \hat{\mathbf{x}}_{1|0}) + \sum_{i=1}^{k} {\mathbf{A_e}}^{k-i}\mathbf{B}(\mathbf{e}^{'}[i] - \mathbf{e}[i])\end{aligned} \tag{78}$$

The limit mean of $\bar{\mathbf{g}}[k+1]$ is:

$$\begin{aligned}&\lim_{k\to\infty} \mathbb{E}[\bar{\mathbf{g}}[k+1]] = \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[\boldsymbol{\upsilon}[k+1]] \\ &= \lim_{k\to\infty} \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[\mathbf{y}^{'}[k+1] - \mathbf{C}\hat{\mathbf{x}}^{'}_{k+1|k} + \mathbf{C}(\hat{\mathbf{x}}^{'}_{k+1|k} \\ &\quad - \hat{\mathbf{x}}_{k+1|k})] \\ &= \lim_{k\to\infty} \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[\boldsymbol{\upsilon}^{'}[k+1]] + \bar{\mathbf{U}}^H\mathbf{KC}\mathbb{E}[{\mathbf{A_e}}^k(\hat{\mathbf{x}}^{'}_{1|0} - \hat{\mathbf{x}}_{1|0}) \\ &\quad + \sum_{i=1}^{k} {\mathbf{A_e}}^{k-i}\mathbf{B}(\mathbf{e}^{'}[i] - \mathbf{e}[i])] \\ &= \lim_{k\to\infty} \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[\boldsymbol{\upsilon}^{'}[k+1]] + \bar{\mathbf{U}}^H\mathbf{KC}\mathbb{E}[{\mathbf{A_e}}^k(\hat{\mathbf{x}}^{'}_{1|0} - \hat{\mathbf{x}}_{1|0})] \\ &\quad + \sum_{i=1}^{k} \bar{\mathbf{U}}^H\mathbf{KC}\mathbb{E}[{\mathbf{A_e}}^{k-i}\mathbf{B}(\mathbf{e}^{'}[i] - \mathbf{e}[i])] \\ &= \mathbf{0}_q\end{aligned} \tag{79}$$

where the first term is the innovation of the virtual system, which has zero mean. The second term will converge to zero because $\mathbf{A_e}$ is stable. The third term is zero because the watermark has zero mean.

The limit covariance of $\bar{\mathbf{g}}[k+1]$ is:

$$\begin{aligned}&\lim_{k\to\infty} \mathrm{Cov}[\bar{\mathbf{g}}[k+1]] \\ &= \lim_{k\to\infty} \bar{\mathbf{U}}^H\mathbf{K}\mathrm{Cov}[\mathbf{y}^{'}[k+1] - \mathbf{C}\hat{\mathbf{x}}^{'}_{k+1|k} + \mathbf{C}(\hat{\mathbf{x}}^{'}_{k+1|k} \\ &\quad - \hat{\mathbf{x}}_{k+1|k})]\mathbf{K}^T\bar{\mathbf{U}} \\ &= \lim_{k\to\infty} \bar{\mathbf{U}}^H\mathbf{K}(\mathrm{Cov}[\boldsymbol{\upsilon}^{'}[k+1]] + \sum_{i=0}^{k} \mathrm{Cov}[\mathbf{C}{\mathbf{A_e}}^i\mathbf{B}\mathbf{e}^{'}[k-i]] \\ &\quad + \sum_{i=0}^{k} \mathrm{Cov}[\mathbf{C}{\mathbf{A_e}}^i\mathbf{B}\mathbf{e}[k-i]])\mathbf{K}^T\bar{\mathbf{U}} \\ &= \mathbf{D} + 2\sum_{i=0}^{\infty} \bar{\mathbf{U}}^H\mathbf{KC}{\mathbf{A_e}}^i\mathbf{B}\boldsymbol{\Sigma}_\mathbf{e}\mathbf{B}^T({\mathbf{A_e}}^T)^i\mathbf{C}^T\mathbf{K}^T\bar{\mathbf{U}}\end{aligned} \tag{80}$$

using the fact that the innovation is independent of the dynamic watermark. Define $\mathbf{X}$ as the solution of the following Lyapunov equation:

$$\mathbf{A_e}\mathbf{X}\mathbf{A_e^T} - \mathbf{X} + \mathbf{B}\boldsymbol{\Sigma}_\mathbf{e}\mathbf{B^T} = \mathbf{0} \tag{81}$$

since $\mathbf{A_e}$ is stable,

$$\mathbf{X} = \sum_{i=0}^{\infty} {\mathbf{A_e}}^i\mathbf{B}\boldsymbol{\Sigma}_\mathbf{e}\mathbf{B}^T({\mathbf{A_e}}^T)^i \tag{82}$$

thus, the covariance of $\bar{\mathbf{g}}[k+1]$ is:

$$\mathrm{Cov}[\bar{\mathbf{g}}[k+1]] = \mathbf{D} + 2\bar{\mathbf{U}}^H\mathbf{K}\mathbf{C}\mathbf{X}\mathbf{C}^T\mathbf{K}^T\bar{\mathbf{U}} \tag{83}$$

The covariance of $\bar{\mathbf{g}}[k+1]$ and $\mathbf{e}[k]$ is:

$$\begin{aligned}
&\mathbb{E}[\bar{\mathbf{g}}[k+1]\mathbf{e}[k]^T] \\
&= \mathbb{E}[\bar{\mathbf{U}}^H\mathbf{K}_{k+1}(\mathbf{C}(\mathbf{x}[k+1-l] - \mathbf{A_d}\hat{\mathbf{x}}_{k|k} - \mathbf{B}\mathbf{h}(\mathbf{z}^k)) \\
&\qquad + \mathbf{n}[k+1-l] - \mathbf{C}\mathbf{B}\mathbf{e}[k])\mathbf{e}[k]^T] \\
&= \bar{\mathbf{U}}^H\mathbf{K}\mathbb{E}[-\mathbf{C}\mathbf{B}\mathbf{e}[k]\mathbf{e}[k]^T] \\
&= -\bar{\mathbf{U}}^H\mathbf{K}\mathbf{C}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{e}
\end{aligned} \tag{84}$$

since other terms are independent of the current dynamic watermark $\mathbf{e}[k]$. Combined with the proofs of mean and covariance completes the proof of equation (47) and (48).

Appendix C
Proof of equation (52)

Under the destabilization attack, the attacked control input:

$$\mathbf{u}_a[k] = \mathbf{u}[k] + \mathbf{A_p}\mathbf{x}[k] \tag{85}$$

the state at $k+1$ if there is no attack is:

$$\mathbf{x}[k+1] = \mathbf{A_d}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{w}[k+1] \tag{86}$$

and denote the state at $k+1$ after attack as:

$$\begin{aligned}
\mathbf{x}_a[k+1] &= \mathbf{A_d}\mathbf{x}[k] + \mathbf{B}\mathbf{u}_a[k] + \mathbf{w}[k+1] \\
&= \mathbf{x}[k+1] + \mathbf{B}\mathbf{A_p}\mathbf{x}[k]
\end{aligned} \tag{87}$$

the measurement at $k+1$ is:

$$\begin{aligned}
\mathbf{z}[k+1] &= \mathbf{C}\mathbf{x}_a[k+1] + \mathbf{n}[k+1] \\
&= \mathbf{y}[k+1] + \mathbf{C}\mathbf{B}\mathbf{A_p}\mathbf{x}[k]
\end{aligned} \tag{88}$$

The innovation after attack is:

$$\begin{aligned}
\boldsymbol{\upsilon}_a[k+1] &= \mathbf{z}[k+1] - \mathbf{C}\hat{\mathbf{x}}_{k+1|k} \\
&= \mathbf{y}[k+1] - \mathbf{C}\hat{\mathbf{x}}_{k+1|k} + \mathbf{C}\mathbf{B}\mathbf{A_p}\mathbf{x}[k]
\end{aligned} \tag{89}$$

so the innovation mean is:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\upsilon}_a[k+1]] &= \mathbb{E}[\mathbf{y}[k+1] - \mathbf{C}\hat{\mathbf{x}}_{k+1|k} + \mathbf{C}\mathbf{B}\mathbf{A_p}\mathbf{x}[k]] \\
&= \mathbf{C}\mathbf{B}\mathbf{A_p}\mathbb{E}[\mathbf{x}[k]] \\
&= \mathbf{C}\mathbf{B}\mathbf{A_p}\hat{\mathbf{x}}_{k|k}
\end{aligned} \tag{90}$$

the first two terms are the innovation without attack which has zeros mean, and by the definition the posterior estiation should be unbiased. The mean of whitened statistic $\bar{\mathbf{g}}[k+1]$ is:

$$\mathbb{E}[\bar{\mathbf{g}}[k+1]] = \bar{\mathbf{U}}^H\mathbf{K}\mathbf{C}\mathbf{B}\mathbf{A_p}\hat{\mathbf{x}}_{k|k} \tag{91}$$

The covariance of the innovation is:

$$\begin{aligned}
&\mathrm{Cov}[\boldsymbol{\upsilon}_a[k+1]] \\
&= \mathrm{Cov}[\boldsymbol{\upsilon}[k+1] + \mathbf{C}\mathbf{B}\mathbf{A_p}(\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})] \\
&= \mathbf{R} + \mathbf{C}\mathbf{B}\mathbf{A_p}\mathbf{P}{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T \\
&\qquad + \mathbb{E}[\boldsymbol{\upsilon}[k+1](\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})^T{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T] \\
&\qquad + \mathbb{E}[\mathbf{C}\mathbf{B}\mathbf{A_p}(\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})\boldsymbol{\upsilon}[k+1]^T]
\end{aligned} \tag{92}$$

where

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{\upsilon}[k+1](\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})^T{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T] \\
&= \mathbb{E}[(\mathbf{C}(\mathbf{x}[k+1] - \hat{\mathbf{x}}_{k+1|k}) + \mathbf{n}[k+1])(\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})^T \\
&\qquad {\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T] \\
&= \mathbf{C}\mathbb{E}[(\mathbf{x}[k+1] - \hat{\mathbf{x}}_{k+1|k})(\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})^T]{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T \\
&= \mathbf{C}\mathbb{E}[(\mathbf{A_d}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{w}[k+1] - \mathbf{A_d}\hat{\mathbf{x}}_{k|k} - \mathbf{B}\mathbf{u}[k]) \\
&\qquad (\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})^T]{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T \\
&= \mathbf{C}\mathbf{A_d}\mathbb{E}[(\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})(\mathbf{x}[k] - \hat{\mathbf{x}}_{k|k})^T]{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T \\
&= \mathbf{C}\mathbf{A_d}\mathbf{P}{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T
\end{aligned} \tag{93}$$

Plug it to the covariance equation:

$$\begin{aligned}
&\mathrm{Cov}[\boldsymbol{\upsilon}_a[k+1]] \\
&= \mathbf{R} + \mathbf{C}\mathbf{B}\mathbf{A_p}\mathbf{P}{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T + \mathbf{C}\mathbf{A_d}\mathbf{P}{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T \\
&\qquad + \mathbf{C}\mathbf{B}\mathbf{A_p}\mathbf{P}{\mathbf{A_p}}^T\mathbf{B}^T\mathbf{C}^T
\end{aligned} \tag{94}$$

Define

$$\mathbf{P_a} = \mathbf{B}\mathbf{A_p}\mathbf{P}{\mathbf{A_p}}^T\mathbf{B}^T + \mathbf{A_d}\mathbf{P}{\mathbf{A_p}}^T\mathbf{B}^T + \mathbf{B}\mathbf{A_p}\mathbf{P}{\mathbf{A_d}}^T \tag{95}$$

then the covariance of $\bar{\mathbf{g}}[k+1]$ is:

$$\mathrm{Cov}[\bar{\mathbf{g}}[k+1]] = \mathbf{D} + \bar{\mathbf{U}}^H\mathbf{K}\mathbf{C}\mathbf{P_a}\mathbf{C}^T\mathbf{K}^T\bar{\mathbf{U}} \tag{96}$$

The covariance of $\bar{\mathbf{g}}[k+1]$ and $\mathbf{e}[k]$ is zero since all terms in $\boldsymbol{\upsilon}[k+1]$ are independent of $\mathbf{e}[k]$. Combined with the proofs of mean and covariance completes the proof of (52).

References

[1] I. O'MALLEY, "U.S. renewable electricity surpassed coal in 2022," *Associated Press*, 2023.

[2] M. Z. Gunduz and R. Das, "Cyber-security on smart grid: Threats and potential solutions," *Computer networks*, vol. 169, p. 107094, 2020.

[3] S. K. Mazumder, A. Kulkarni, S. Sahoo, F. Blaabjerg, H. A. Mantooth, J. C. Balda, Y. Zhao, J. A. Ramos-Ruiz, P. N. Enjeti, P. Kumar *et al.*, "A review of current research trends in power-electronic innovations in cyber–physical systems," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 5, pp. 5146–5163, 2021.

[4] J. C. Balda, A. Mantooth, R. Blum, and P. Tenti, "Cybersecurity and power electronics: Addressing the security vulnerabilities of the internet of things," *IEEE Power Electronics Magazine*, vol. 4, no. 4, pp. 37–43, 2017.

[5] A. Walker, J. Desai, D. Saleem, and T. Gunda, "Cybersecurity in photovoltaic plant operations," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2021.

[6] X. Liu, M. Shahidehpour, Y. Cao, L. Wu, W. Wei, and X. Liu, "Microgrid risk analysis considering the impact of cyber attacks on solar pv and ess control systems," *IEEE transactions on smart grid*, vol. 8, no. 3, pp. 1330–1339, 2016.

[7] A. Teymouri, A. Mehrizi-Sani, and C.-C. Liu, "Cyber security risk assessment of solar pv units with reactive power capability," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2018, pp. 2872–2877.

[8] S. Nath, I. Akingeneye, J. Wu, and Z. Han, "Quickest detection of false data injection attacks in smart grid with dynamic models," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 10, no. 1, pp. 1292–1302, 2019.

[9] I. Akingeneye and J. Wu, "Pmu-assisted bad data detection in power systems," in *2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*. IEEE, 2018, pp. 1–5.

[10] D. M. Shilay, K. G. Lorey, T. Weiz, T. Lovetty, and Y. Cheng, "Catching anomalous distributed photovoltaics: An edge-based multi-modal anomaly detection," *arXiv preprint arXiv:1709.08830*, 2017.

[11] K. G. Lore, D. M. Shila, and L. Ren, "Detecting data integrity attacks on correlated solar farms using multi-layer data driven algorithm," in *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2018, pp. 1–9.

[12] M. Mohammadpourfard, Y. Weng, I. Genc, and T. Kim, "An accurate false data injection attack (fdia) detection in renewable-rich power grids," in *2022 10th Workshop on Modelling and Simulation of Cyber-Physical Energy Systems (MSCPES)*. IEEE, 2022, pp. 1–5.

[13] A. Moradzadeh, M. Mohammadpourfard, I. Genc, Ş. S. Şeker, and B. Mohammadi-Ivatloo, "Deep learning-based cyber resilient dynamic line rating forecasting," *International Journal of Electrical Power & Energy Systems*, vol. 142, p. 108257, 2022.

[14] Q. Li, F. Li, J. Zhang, J. Ye, W. Song, and A. Mantooth, "Data-driven cyberattack detection for photovoltaic (pv) systems through analyzing micro-pmu data," in *2020 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2020, pp. 431–436.

[15] J. Zhang, L. Guo, and J. Ye, "Hardware-in-the-loop testbed for cyber-physical security of photovoltaic farms," in *2021 IEEE 12th International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*. IEEE, 2021, pp. 1–7.

[16] S. Nath and J. Wu, "Quickest change point detection with multiple postchange models," *Sequential Analysis*, vol. 39, no. 4, pp. 543–562, 2020.

[17] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2013.

[18] R. Tunga, C. Murguia, and J. Ruths, "Tuning windowed chi-squared detectors for sensor attacks," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 1752–1757.

[19] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th annual Allerton conference on communication, control, and computing (Allerton)*. IEEE, 2009, pp. 911–918.

[20] K. Manandhar, X. Cao, F. Hu, and Y. Liu, "Detection of faults and attacks including false data injection attack in smart grid using kalman filter," *IEEE transactions on control of network systems*, vol. 1, no. 4, pp. 370–379, 2014.

[21] B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber–physical systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, 2016.

[22] J. Ramos-Ruiz, J. Kim, W.-H. Ko, T. Huang, P. Enjeti, P. Kumar, and L. Xie, "An active detection scheme for cyber attacks on grid-tied pv systems," in *2020 IEEE CyberPELS (CyberPELS)*. IEEE, 2020, pp. 1–6.

[23] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Dynamic watermarking for general lti systems," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 1834–1839.

[24] M. Porter, P. Hespanhol, A. Aswani, M. Johnson-Roberson, and R. Vasudevan, "Detecting generalized replay attacks via time-varying dynamic watermarking," *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3502–3517, 2020.

[25] J. Panchal, B. Wen, and R. Burgos, "Power electronics based self-monitoring and diagnosing for photovoltaics systems," in *2021 IEEE 22nd Workshop on Control and Modelling of Power Electronics (COMPEL)*. IEEE, 2021, pp. 1–8.

[26] K. Dhibi, M. Mansouri, K. Bouzrara, H. Nounou, and M. Nounou, "An enhanced ensemble learning-based fault detection and diagnosis for grid-connected pv systems," *IEEE Access*, vol. 9, pp. 155 622–155 633, 2021.

[27] X. Jiao, X. Li, T. Yang, Y. Yang, and W. Xiao, "A novel fault diagnosis scheme for pv plants based on real-time system state identification," *IEEE Journal of Photovoltaics*, 2023.

[28] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, no. 6, pp. 601–611, 1976.

[29] F. Gustafsson and F. Gustafsson, *Adaptive filtering and change detection*. Citeseer, 2000, vol. 1.

[30] I. Akingeneye and J. Wu, "Low latency detection of sparse false data injections in smart grids," *IEEE Access*, vol. 6, pp. 58 564–58 573, 2018.

[31] M. Basseville, "Detecting changes in signals and systems—a survey," *Automatica*, vol. 24, no. 3, pp. 309–326, 1988.

[32] S. Nath and J. Wu, "Bayesian quickest change point detection with multiple candidates of post-change models," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 51–55.

[33] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," in *2016 IEEE Conference on Control Applications (CCA)*. IEEE, 2016, pp. 474–480.

[34] L. Chen, S. Li, and X. Wang, "Quickest fault detection in photovoltaic systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1835–1847, 2016.

[35] M. Izbicki, S. Amini, C. R. Shelton, and H. Mohsenian-Rad, "Identification of destabilizing attacks in power systems," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 3424–3429.

[36] J. Tang, J. Song, and A. Gupta, "A dynamic watermarking algorithm for finite markov decision problems," *arXiv preprint arXiv:2111.04952*, 2021.