An Online Detection Framework for Cyber Attacks on Automatic Generation Control

Tong Huang , *Student Member, IEEE*, Bharadwaj Satchidanandan , P. R. Kumar, *Life Fellow, IEEE*, and Le Xie , *Senior Member, IEEE*

Abstract—We propose an online framework to detect cyberattacks on automatic generation control. A cyberattack detection algorithm is designed based on the approach of dynamic watermarking. The detection algorithm provides a theoretical guarantee of detection of cyberattacks launched by sophisticated attackers possessing extensive knowledge of the physical and statistical models of targeted power systems. The proposed framework is practically implementable, as it needs no hardware update on generation units. The efficacy of the proposed framework is validated in both four-area system and 140-bus system.

Index Terms—Dynamic watermarking, cyber-physical security (CPS), automatic generation control (AGC).

I. INTRODUCTION

THE role of Automatic Generation Control (AGC) in large power systems is indispensable. It maintains nominal frequency while minimizing generation costs. The operation of the AGC involves close interaction between the cyber and the physical layers. By tracking Area Control Error (ACE) deviation collected from distributed sensors, the power outputs of generators are modified via AGC to balance random fluctuation of loads, and the electric grid frequency is thereby maintained within a tight range around the nominal value (50/60 Hz). However, due to the consequent tight coupling between the cyber and physical layers, there arises a vulnerability in that both grid stability and security can be compromised by malicious attacks on the cyber layer for sensing. Rather than compromising the strongly secured cyber layers of the control centers, cyber attacks on distributed measurements feeding the AGC might in fact significantly disrupt the operational goals of the power system [1].

There have been several efforts at examining the potential mechanisms by which such cyber attacks on AGCs can be car-

ried out and their negative impacts on the system operation. For example, as described in [2], several attempts for cyber attacks on AGCs, namely, scaling, ramp, pulse, and random attacks, may compromise both the physical system stability and the electricity market operation. Experiments based on CPS testbeds suggest that the corrupted measurements feeding the AGC might bring power systems to under-frequency condition and cause unnecessary load shedding [3], [4]. By replacing the original measurements with an "optimal attack sequence", the malicious attackers can disrupt the system frequency in the shortest time without triggering certain pre-defined data quality alarms [1]. Besides cyber attacks on AGC, potential risks can also be posed from the load side: adversaries may be able to trip targeted generators by manipulating the controller parameters of the loads offering emulated inertia control services [5], [6]. This paper focuses on the detection of cyber attacks on AGC.

All of the above attack strategies on AGC are based on the assumption that the cyber layer of the AGC transporting the physical measurements is vulnerable to attacks, so that a malicious adversary can manipulate these measurements. Unfortunately, this assumption is validated by several recent real-world incidents. Examples include computer viruses such as Dragonfly [7] and Stuxnet [8] targeting Industrial Control Systems (ICS). Therefore, although no real-world attack specifically targeting the AGC has been reported thus far, the aforementioned attack strategies on AGC are more than theoretical concerns. As grid operation becomes more and more data-dependent, it is imperative to prepare the operators with an online defense mechanism against all possible cyber attacks on AGCs.

There have been several detection techniques for cyber attacks on AGCs. In [2], cyber attacks following predefined attack strategies are detected by checking the statistical and temporal characterization of area control errors (ACE). In [9], a statistical model learned from frequency and tie-line flow measurements is exploited to predict their short-term values. Measurements in the vicinity of their corresponding predictions are tagged as normal measurements. Otherwise, alarms are triggered. In [1], the compromised tie-line flow measurements are detected by capturing the discrepancy between the meter readings of frequency deviation and its predicted value based on reported tie-line flow measurements and an identified linear-regression model.

Also, DC state estimation (SE) is modified to be executed every AGC cycle and serves as an additional layer for data purification in [1].

Although the aforementioned approaches increase the attack costs to some extent, the measurements feeding the AGC may still be compromised by an attacker equipped with the following capabilities. First, the malicious adversaries are not constrained to follow the prescribed attack templates in order to cause significant impact on the grid [1]. Although the anomaly detection engine proposed in [2] is capable of identifying the predefined attack templates, there is no theoretical guarantee that the proposed algorithm can detect arbitrary cyber attacks. Second, extensive information on the system model might be exposed to the adversary. There are two ways by which a malicious adversary can obtain information about the power system model: 1) The detailed physical model may be directly leaked to the attacker via disgruntled employees or malicious insiders [10]; 2) The statistical model of the power system can be learned using mathematical tools based on the leaked system operating data. The attackers in the former case can bypass the SE-based detection algorithm by conducting "unobservable attacks" described in [11] or by conducting the packet-reordering integrity attack reported in [12], whereas the adversaries in the latter case can tamper with the measurements without triggering the alarm defined in [9] by replacing the actual measurement sequence with a different sequence that still conforms to the learned statistical model [13]. Besides, the authors of [1] exclude the attacks on frequency sensors from their framework. Therefore, a subtle but malicious distortion of frequency measurements based on the physical/statistical model of the power system is not likely to be detected by the algorithm proposed in [1].

In this paper, we introduce a first-of-its-kind online detection framework of false data injection attacks in power systems. The recent dynamic watermarking technique [13], [14] is employed in the framework and serves as the core algorithm to detect any tampered measurements feeding the AGC. Through deliberately superimposing a private signal of small magnitude upon the control commands sent by the AGC, we "watermark" the measurements feeding the AGC with certain indelible characteristics [13], by which cyber attacks on the AGC can be identified. To the best of the authors' knowledge, this is the first time that the dynamic watermarking technique has been applied to address cyber-security issues in power systems. The proposed framework has the following advantages. 1) The detection algorithm used with the dynamic watermarking is theoretically rigorous and ensures that any manipulation of the measurements feeding the AGC can be detected regardless of the attack strategy that the attackers follow, as long as the controlled generators can execute commands from AGC honestly; 2) the algorithm can be used when attackers possess detailed information of the physical/statistical models of the power system; 3) the proposed framework is practically implementable, as it needs no hardware update on generation units.

The rest of this paper is organized as follows. Section II formulates the problem of detection of cyber attacks by mathematically describing a system equipped with AGC and by presenting typical attack models; Section III presents the dynamic watermarking-based detection algorithm in the context of AGC; Section IV validates the efficacy of the proposed algorithm via an illustrative example; Section V concludes the paper.

II. PROBLEM FORMULATION

In this section, a power system equipped with multiple AGCs is described mathematically, and typical attack templates are presented.

A. The Model of a Multi-Area Power System Without AGC

The dynamics of a multi-area power system in the vicinity of an operating condition can be described approximately by a continuous state-space model [15]:

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t) + \boldsymbol{\gamma}'(t), \tag{1a}$$

$$\boldsymbol{y}(t) = C\boldsymbol{x}(t) + \boldsymbol{n}'(t), \tag{1b}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^{n' \times 1}$, $\boldsymbol{u}(t) \in \mathbb{R}^{d \times 1}$ and $\boldsymbol{y}(t) \in \mathbb{R}^{m \times 1}$ are states, inputs and measurements vectors in the time instant t, respectively, and the matrices A, B and C are system parameters of appropriate dimensions. Above $\boldsymbol{\gamma}'(t) \sim \mathcal{N}(0, Q')$ and $\boldsymbol{n}'(t) \sim \mathcal{N}(0, R')$ denote the white process noise and the measurement noise respectively that are independent of each other (A more mathematical description would entail stochastic differential equations). Suppose that there are r control areas. Then, the measurement vector $\boldsymbol{y}(t)$ in (1) can be reorganized as $\boldsymbol{y}(t) = [\boldsymbol{y}_1(t)^T \ \boldsymbol{y}_2(t)^T \ \cdots \ \boldsymbol{y}_i(t)^T \ \cdots \ \boldsymbol{y}_r(t)^T]^T$, where $(\cdot)^T$ is the transpose operation, and $\boldsymbol{y}_i(t)$ is a column vector incorporating all tie-line flow deviations $\boldsymbol{p}_{ti}(t)$, as well as the frequency deviation $\omega_i(t)$ in the control area i, i.e.,

$$\boldsymbol{y}_i(t) = \begin{bmatrix} \boldsymbol{p}_{ti}(t)^T & \omega_i(t) \end{bmatrix}^T. \tag{2}$$

Similarly, the variables in $\boldsymbol{u}(t)$ can be grouped area-wise into $\boldsymbol{u}(t) = [\boldsymbol{u}_1(t)^T \ \boldsymbol{u}_2(t)^T \ \cdots \ \boldsymbol{u}_i(t)^T \ \cdots \ \boldsymbol{u}_r(t)^T]^T$, where the column vector $\boldsymbol{u}_i(t)$ includes the load reference setpoints $\boldsymbol{p}_{si}(t) \in \mathbb{R}^{d' \times 1}$ of all generators participating in AGC in the area i, as well as local load fluctuation $\boldsymbol{p}_{\text{load}i}(t) + j\boldsymbol{q}_{\text{load}i}(t)$ at time instant t, i.e.,

$$\boldsymbol{u}_i(t) = \begin{bmatrix} \boldsymbol{p}_{si}(t)^T & \boldsymbol{u}_{\text{load}i}(t)^T \end{bmatrix}^T \tag{3}$$

where $\boldsymbol{u}_{\text{load}i} = [\boldsymbol{p}_{\text{load}i}(t)^T \ \boldsymbol{q}_{\text{load}i}(t)^T]^T$.

B. The Model of a Multi-Area System Regulated by AGC

From a system-theoretic perspective, the AGC can be regarded as a multi-variable feedback loop added to the plant described in (1). In order to achieve independent regulation for the local tie-line flows and frequency, the Balancing Authority in one area only actuates the local generators participating in AGC without interference from generators in other areas. Therefore, the multi-area control policy can be decentralized area-wise as

$$\begin{aligned} \mathbf{u}[t] &= \boldsymbol{f}(\boldsymbol{y}^t) \\ &= \begin{bmatrix} \boldsymbol{f}_1(\boldsymbol{y}_1^t)^T & \boldsymbol{f}_2(\boldsymbol{y}_2^t)^T & \cdots & \boldsymbol{f}_i(\boldsymbol{y}_i^t)^T & \cdots & \boldsymbol{f}_r(\boldsymbol{y}_r^t)^T \end{bmatrix}^T \end{aligned} \tag{4}$$

where $\boldsymbol{y}_i^t$ is the telemetered measurement sequence up to time t at area i. To elaborate on the control policy, suppose that there are ψ local generation units in the AGC and ϕ measurements in area i, then the control policy of AGC $\boldsymbol{f}_i(\cdot) : \mathbb{R}^\phi \to \mathbb{R}^\psi$ consists of the following operations between two successive economic dispatches:
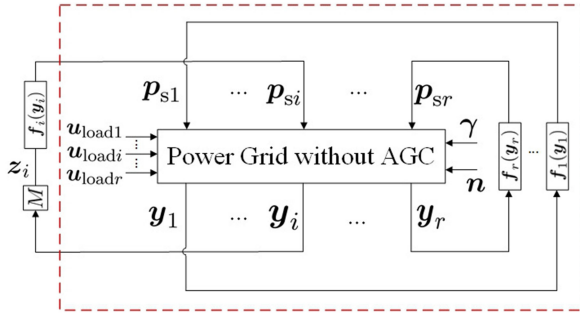
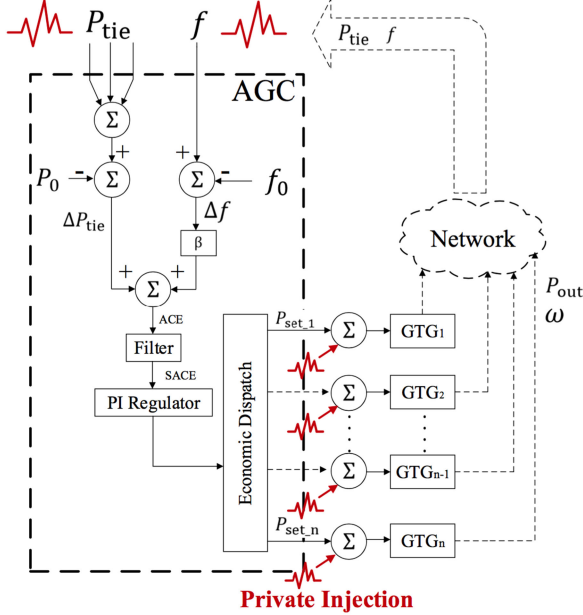Fig. 1. A multi-area power system with AGC systems.

Fig. 2. Location of private injection in a simplified functional diagram of AGC.

1) Area control error (ACE) is calculated from the teleme-tered tie-line flows and frequency measurements sampled every two to four seconds as

$$ACE_i = \sum_{s=1}^{\phi} p_{ti,s} + \beta_i \omega_i,$$

where the adjustable parameter β_i is a bias factor.
2) The ACE is smoothed by passing it through a low-order filter in order to mitigate the fatigue of generation control devices, e.g., turbine valves and governor motors [16].
3) At the balancing authority, a control command is computed from the ACE according to the control policy reported in [17], and is executed every two to four seconds [16], [10]. Denote by $\kappa_i \tau$ the time period between two consecutive commands.
4) The control command computed by AGC is sent to the ψ local generation units and its magnitude for each controlled generator is proportional to the coefficient updated by the economic dispatch algorithm [18], [19].

The above procedure (also summarized in Fig. 2) indicates that only the measurements at the chosen sample instants contribute to the computation of the control commands sent by the AGC at area i. The sequence $\boldsymbol{y}_i^t$ formed by these measurements is denoted by

$$\boldsymbol{y}_i^t := \left\{ \boldsymbol{y}_i(0), \boldsymbol{y}_i(\kappa_i \tau), \ldots, \boldsymbol{y}_i \left(\left\lfloor \frac{t}{\kappa_i \tau} \right\rfloor \kappa_i \tau \right) \right\} \quad (5)$$

where $\lfloor \cdot \rfloor$ is the floor function. The above control policy yields the load reference setpoints $\boldsymbol{p}_{si}(t)$, so that

$$\boldsymbol{p}_{si}(t) = \boldsymbol{f}_i(\boldsymbol{y}_i^t) \quad \forall i \in \{1, 2, \ldots, r\}. \quad (6)$$

The above equation couples the physical infrastructure (generation units) and the cyber layer (control centers) together. In summary, (1), (4) and (6) constitute a hybrid model for a multi-area power system regulated by AGCs.

Note that a commercial-level AGC includes more functions, which are assumed to be included into the control law $\boldsymbol{f}_i(\cdot)$. Fig. 2 shows a simplified version of a realistic AGC.

C. Discretization of the Hybrid AGC Model

Suppose that the time period between two consecutive control commands of AGC in each area is an integer multiple of a sampling time τ, namely, κ_i is assumed to be an integer. Then the continuous-time state space model (1) can be discretized at τ using the approach reported in [20]. For the sake of convenience, the discrete state-space model is denoted as $\mathcal{M}_d''$. Similarly, the AGC control policies in area can also be sampled at τ. Denote the discrete control policies by $\boldsymbol{f}_{di}(\cdot)$ for all $i \in \{1, 2, \ldots, r\}$. It is worth noting that all areas are sampled with the same interval τ, and the AGC in area i sends control signals only after every $\kappa_i \tau$ seconds, for $i \in \{1, 2, \ldots, r\}$.

For the control area i, we temporarily open its AGC feedback loop and keep the AGC loops in other areas j connected, for $j \in \{1, 2, \ldots, r\}$ and $j \neq i$. As shown in Fig. 1, we focus on modeling the open-loop behavior of the system for area i in terms of its inputs, i.e., the setpoints $\boldsymbol{p}_{si}$ of the controlled generators in the area i, and all load fluctuations $\boldsymbol{u}_{\text{load}j}$ for all $j \in \{1, 2, \ldots, r\}$, and its outputs, i.e., all tie-line flow deviations $\boldsymbol{p}_{ti}$ and frequency deviations ω_i in (2). As is standard in linear control theory [21], the discrete model of the aforementioned open-loop system can be obtained by interconnecting the entire system model $\mathcal{M}_d''$ and the discrete AGC control policies $\boldsymbol{f}_{dj}(\cdot)$, where $j \in \{1, 2, \ldots, r\}$ and $j \neq i$. Denote the resulting interconnected state-space model for area i by $\mathcal{M}_{di}'$. It is worth noting that the state variables of $\mathcal{M}_{di}'$ include all state variables in both state-space model $\mathcal{M}_d''$ and discrete control policies $\boldsymbol{f}_{dj}$, where $j \in \{1, 2, \ldots, r\}$ and $j \neq i$. We specify setpoint $\boldsymbol{p}_{si}$ as the control inputs of system $\mathcal{M}_{di}'$, and further assume $\mathcal{M}_{di}'$ is stabilizable [20]. Finally, the discrete state-space model $\mathcal{M}_{di}'$ can be minimally realized by a controllable and observable model $\mathcal{M}_{di}$ with reduced order [20], namely,

$$\boldsymbol{x}_{di}(k+1) = A_{di}\boldsymbol{x}_{di}(k) + B_{di}^{\text{ref}}\boldsymbol{p}_{si}(k) \\ + B_{di}^{\text{load}}\boldsymbol{u}_{\text{load}}(k) + \boldsymbol{\gamma}(k+1) \quad (7a)$$

$$\boldsymbol{y}_i(k) = C_{di}\boldsymbol{x}_{di}(k) + \boldsymbol{n}(k) \quad (7b)$$

where $\boldsymbol{x}_{\mathrm{d}i} \in \mathbb{R}^{n \times 1}$ collects all state variables in the reduced-order model $\mathcal{M}_{\mathrm{d}i}$ and $\boldsymbol{u}_{\mathrm{load}}(k) = [\boldsymbol{u}_{\mathrm{load}1}^T \ \boldsymbol{u}_{\mathrm{load}2}^T \ \cdots \ \boldsymbol{u}_{\mathrm{load}r}^T]^T$. Vector $\boldsymbol{\gamma}(t) \sim \mathcal{N}(0, Q)$ and $\boldsymbol{n}(t) \sim \mathcal{N}(0, R)$ are the white process and measurement noises, where R is positive definite. We assume that the rank of matrix $C_{\mathrm{d}i} B_{\mathrm{d}i}^{\mathrm{ref}}$ equals ϕ, which is the number of rows of $C_{\mathrm{d}i}$.

D. Cyber Attack Models and Their Impacts

Due to the close interaction between the AGC and the generation units indicated by (6), the adversary can compromise the physical layer of the power system by distorting the measurements $\boldsymbol{y}^t$. Denote by $\boldsymbol{z}^t$ the measurements *reported* by the sensors. The sensors are supposed to report the actual value measured, i.e., they are supposed to report truthfully with $\boldsymbol{z}^t = \boldsymbol{y}^t$. However, an adversarial sensor might declare values that are different from the actual measurements, so that $\boldsymbol{z}^t \neq \boldsymbol{y}^t$. The purpose of this paper is to detect the inconsistency between the actual and the reported measurements caused deliberately by the attacker. The attackers are assumed to be able to manipulate the distributed sensors feeding into AGC, i.e., frequency and tie-line flow measurements. Before describing the remedy for the problem, we present three typical attack templates.

1) Replay Attack: Before the attack, the adversary records the measurements during normal operating condition for some duration. During the attack, the actual measurements observed by the adversarial sensors are replaced by the recorded measurements and reported to the control center [22].

2) Noise-Injection Attack: Under this attack model, the adversarial sensors add a bounded random value to the actual measurement and then report it to the control center.

3) Destabilization Attack: In a destabilization attack, the compromised sensors of the AGC in area i report a sequence $\{\boldsymbol{z}_i\}$ which is a filtered version of the actual measurement sequence $\{\boldsymbol{y}_i\}$. If M denotes such a filter, the attack consists of inserting the filter M to the system model, with M so chosen such that the original system becomes unstable. It is worth noting that the output sequence $\boldsymbol{z}_i$ of a malicious filter M can be obtained through a simple tuning procedure, even without any information on the system model, as will be described in Section IV-A5.

Note that the attackers are not limited to follow any attack templates, in their attempt to bring harm to power systems. Correspondingly, a defense method should be designed not only for detecting the three types of attacks defined above, but also for securing AGC from any manipulation on the distributed measurements feeding into AGC.

III. DYNAMIC WATERMARKING-BASED DEFENSE

In this section, we apply the approach of dynamic watermarking reported in [13], [14], [23] to secure the distributed measurements feeding AGC in power systems. The fundamental idea of Dynamic Watermarking is as follows. The actuators (generation units in this case) superimpose on the control policy-specified input, a "small" random signal chosen according to a certain probability distribution. While this probability distribution is made public, so that even the adversary knows it, the actual

realization of the random signal is known only to that particular generation unit, and it doesn't reveal that to any other party. For this reason, the random signal is also called the private excitation of the generators. In such a scenario, the honest sensors and the malicious sensors are distinguished by the following fact: the truthful measurements reported by the honest sensors exhibit certain expected statistical properties that are relevant to the statistics of the private excitation, whereas, as shown in [13], [14], measurements reported by the malicious sensors, if excessively distorted, do not exhibit these properties. Therefore, by subjecting the reported measurements to certain tests for these statistical properties, malicious activity in the system can be detected.

In this paper, we will demonstrate the application of this approach in the context of power systems. For control area i, an independent and identically distributed (i.i.d.) private excitation $\{\boldsymbol{e}_i(k)\}$ is superimposed on the control inputs $\{\boldsymbol{p}_{si}(k)\}$ [14]. Consequently, the input applied at time k is

$$\boldsymbol{p}_{si}(k) = \boldsymbol{f}_i(\boldsymbol{y}_i^k) + \boldsymbol{e}_i(k), \tag{8}$$

where $\boldsymbol{e}_i(k) \sim \mathcal{N}(0, \sigma_e^2 I)$. It is worth noting that (8) can be implemented by modifying the AGC software at the balancing authorities without any hardware updates on the generation units. With the private injection $\{\boldsymbol{e}_i(k)\}$, any attempt to distort the measurements fed to AGC will be detected by subjecting the reported measurements to the two tests [14] described below. A detailed proof for this conclusion can be found in [14].

A. Two Indicators of Dynamic Watermarking

Given the input sequence $\boldsymbol{u}_i$ and measurement sequence $\boldsymbol{y}_i$ of the discrete system (7) up to the kth unit of time, the system state $\boldsymbol{x}_{\mathrm{d}i}(k|k)$ can be estimated by the Kalman filter as

$$\boldsymbol{x}_{\mathrm{d}i}(k + 1|k) = A_{\mathrm{d}i}(I - L_{\mathrm{d}i} C_{\mathrm{d}i})\boldsymbol{x}_{\mathrm{d}i}(k|k-1) +$$
$$\begin{bmatrix} B_{\mathrm{d}i}^{\mathrm{ref}} \ B_{\mathrm{d}i}^{\mathrm{load}} \ A_{\mathrm{d}i} L_{\mathrm{d}i} \end{bmatrix} \begin{bmatrix} \boldsymbol{p}_{si}(k) \\ \boldsymbol{u}_{\mathrm{load}}(k) \\ \boldsymbol{y}_i(k) \end{bmatrix}, \tag{9a}$$

$$\boldsymbol{x}_{\mathrm{d}i}(k|k) = (I - L_{\mathrm{d}i} C_{\mathrm{d}i})\boldsymbol{x}_{\mathrm{d}i}(k|k-1) + L_{\mathrm{d}i} \boldsymbol{y}_i(k), \tag{9b}$$

where $L_{\mathrm{d}i}$ is the steady-state Kalman filtering gain given by

$$L_{\mathrm{d}i} = PC_{\mathrm{d}i}^T(C_{\mathrm{d}i} PC_{\mathrm{d}i}^T + R)^{-1}. \tag{10}$$

In the above, P is obtained as the unique positive definite solution of the Algebraic Riccati Equation [24]. We define

$$\boldsymbol{\zeta}_k := \boldsymbol{x}_{\mathrm{d}i}(k|k) - A_{\mathrm{d}i}\boldsymbol{x}_{\mathrm{d}i}(k-1|k-1) - B_{\mathrm{d}i}^{\mathrm{ref}}\boldsymbol{f}_i(\boldsymbol{z}_i^{k-1})$$
$$- B_{\mathrm{d}i}^{\mathrm{ref}}\boldsymbol{e}(k-1) - B_{\mathrm{d}i}^{\mathrm{load}}\boldsymbol{u}_{\mathrm{load}}. \tag{11}$$

1) Test 1: Check if

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} \boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^T = L_{\mathrm{d}i} \Sigma_i L_{\mathrm{d}i}^T, \tag{12}$$

where

$$\Sigma_i := C_{\mathrm{d}i} PC_{\mathrm{d}i}^T + R. \tag{13}$$

Correspondingly, we choose a time window T and define an indicator matrix W by

$$W(T) := \frac{1}{T} \sum_{k=1}^{T} \boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^T - L_{\mathrm{d}i} \Sigma_i L_{\mathrm{d}i}^T. \tag{14}$$

2) Test 2: Check if

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} \boldsymbol{e}(k-1) \boldsymbol{\zeta}_k^T = 0. \tag{15}$$

As before, we define

$$V(T) := \frac{1}{T} \sum_{k=1}^{T} \boldsymbol{e}(k-1) \boldsymbol{\zeta}_k^T. \tag{16}$$

This measure can be calculated by the system operators. The reported measurements of interest, $\{\boldsymbol{z}_i(k)\}$, will pass both tests if $\boldsymbol{z}_i(k) \equiv \boldsymbol{y}_i(k)$ for all k; if the sensors distort the measurements beyond adding a zero-power signal, then, as shown in [14], at least one of the above tests will fail. While tests (12) and (15) are asymptotic in nature, they can be converted to statistical tests that can be performed in finite time. For example, we expect much bigger entries in W or V during cyber attacks, than their counterparts when no attack happens. This leads naturally to a threshold test for detecting malicious distortion.

B. Online Algorithm for Detection of Cyber Attacks

The computation of the aforementioned indicators requires a sequence of reported measurements $\{\boldsymbol{z}_i\}$, private injections $\{\boldsymbol{e}_i\}$, load fluctuations of the whole grid $\{\boldsymbol{u}_{\mathrm{load}}\}$ and AGC command signals $\{\boldsymbol{f}_i(\boldsymbol{z}_i^{k-1})\}$ over a period of time. Therefore, in order to check whether the reported measurements pass the two tests (12), (15), the generation unit processes a block of $\{\boldsymbol{z}_i\}$, $\{\boldsymbol{e}_i\}$, $\{\boldsymbol{u}_{\mathrm{load}}\}$ and $\{\boldsymbol{f}_i(\boldsymbol{z}_i^{k-1})\}$ within a time window T. Suppose that each block of the above sequences includes T samples. Then, up to time $t = j \times T \times \kappa_i \tau$, we will have j blocks of above sequences. The jth block of above sequences in area i are denoted by $\boldsymbol{z}_i^{\mathrm{BL}j}$, $\boldsymbol{e}_i^j$, $\boldsymbol{u}_{\mathrm{load}}^j$ and $\boldsymbol{f}_i^j$, respectively:

$$\boldsymbol{z}_i^{\mathrm{BL}j} := \{\boldsymbol{z}_i((j-1)T), \boldsymbol{z}_i((j-1)T+1), \dots, \boldsymbol{z}_i(jT)\},$$

$$\boldsymbol{e}_i^j := \{\boldsymbol{e}_i((j-1)T), \boldsymbol{e}_i((j-1)T+1), \dots, \boldsymbol{e}_i(jT)\},$$

$$\boldsymbol{u}_{\mathrm{load}}^j := \{\boldsymbol{u}_{\mathrm{load}}((j-1)T), \boldsymbol{u}_{\mathrm{load}}((j-1)T+1), \dots, \boldsymbol{u}_{\mathrm{load}}(jT)\},$$

and

$$\boldsymbol{f}_i^j := \{\boldsymbol{f}_i(\boldsymbol{z}_i^{(j-1)T-1}), \boldsymbol{f}_i(\boldsymbol{z}_i^{(j-1)T}), \dots, \boldsymbol{f}_i(\boldsymbol{z}_i^{jT-1})\}.$$

In terms of online application, let $W^j = [w_{g,h}^j]$ and $V^j = [v_{g,h}^j]$ be W and V calculated within the jth time window, respectively. Then the indicator scalars ξ_1^j and ξ_2^j are defined as follows

$$\xi_1^j := |\mathrm{tr}(W^j)| \tag{17a}$$

$$\xi_2^j := \sqrt{\sum_{g=1}^{d'} \sum_{h=1}^{n} (v_{g,h}^j)^2} \tag{17b}$$

where $\mathrm{tr}(\cdot)$ is the trace operator. As mentioned in (2) and (7), d' is the number of the controlled generators in AGC of area i and n is the order of the reduced-order model in (7). Finally, we expect $\xi_1^j \geq \eta_1$ or $\xi_2^j \geq \eta_2$, if attacks are launched in the jth time window, where η_1 and η_2 are pre-defined thresholds. The thresholds η_1 and η_2 can be obtained from the following training procedure:

1) based on (14) and (16), first compute $W^\infty = W(T_\infty)$ and $V^\infty = V(T_\infty)$ under normal operating condition, where T_∞ is a large integer that is set to 1800 in this paper;
2) obtain the general indicators ξ_1^∞ and ξ_2^∞ under a normal condition by (17);
3) the thresholds η_1 and η_2 are calculated by

$$\eta_1 = \kappa' \xi_1^\infty \quad \eta_2 = \kappa' \xi_2^\infty \tag{18}$$

where κ' is an empirically adjustable parameter.

The detection thresholds η_1, η_2 can also be determined using the Neyman-Pearson criterion based on the maximum tolerable false alarm rate, which is shown in Appendix A. Algorithm 2 specifies the subroutine for computing the two indicators ξ_1^j and ξ_2^j for the jth block of measurements.

For area i, private signals $\boldsymbol{e}_i$ are superimposed upon the AGC commands according to (8) and Fig. 2. Then Algorithm 1 enables the balancing authority of area i to detect cyber attacks on the measurements feeding the AGC. Once attacks in area i are detected, the balancing authority stops sending commands to the generators in the AGC. Similarly, attacks to other areas can be detected by the corresponding balancing authorities similarly equipped with Algorithm 1. Additionally, it is worth emphasizing that Fig. 2 is a *simplified* functional diagram of AGC, where the optimal power setpoints are the actual outputs of the simplified AGC. In the proposed method, the private excitations $\boldsymbol{f}_i(\boldsymbol{y}_i^k)$ are supposed to be superimposed upon the actual outputs of AGC, which is not necessary to be the calculated optimal power setpoint in a realistic AGC.

After a cyber attack is detected by the proposed framework, the AGC should be deactivated. It is worth noting that an efficient procedure for finding malicious sensors should be initiated after deactivating the AGC. Such a procedure may include dispatching a panel to investigate the distributed measurements after an alarm. Also, the procedure is required to correct the malicious sensors quickly. This requirement is achievable due to the limited number of the distributed measurements feeding to AGC. After clearing the cyberattacks, the AGC should be back to service. Therefore, the AGC is actually absent only for a short period of time, instead of permanently out of service.

One might wonder if the temporary absence of AGC significantly impacts the system frequency. The answer is that the temporary absence of AGC should not be a big concern, as the AGC is allowed to be deactivated in real-world system

---

Algorithm 1: Online Algorithm for Detection of Cyber Attack.

1: $H \leftarrow L_{\mathrm{d}i}\Sigma_i L_{\mathrm{d}i}^T; j \leftarrow 1$
2: **while** $k = 1, 2, \cdots,$ **do**
3:     **if** $k \geq jT$ **then**
4:         Obtain the sequence $z_i^{\mathrm{BL}j}, e_i^j, u_{\mathrm{load}}^j, f_i^j$;
5:         Compute $x_{\mathrm{e}} := \{x(k'|k')\}$ by (6) and (9)
        for all $k' = (j-1)T, (j-1)T+1, \ldots, jT$;
6:         $\xi_1^j, \xi_2^j \leftarrow$ Indicators $(x_{\mathrm{e}}^j, e_i^j, u_{\mathrm{load}}^j, f_i^j,$
        $j, H)$;
7:         $j \leftarrow j+1$
8:         **if** $\xi_1 \geq \eta_1 \vee \xi_2 \geq \eta_2$ **then**
9:             Claim attacks and stop sending commands
            to the generators on AGC;
10:         **end if**
11:     **end if**
12: **end while**

Algorithm 2: Computation of ξ_1^j and ξ_2^j at the jth block

1: **function** Indicators $(x_{\mathrm{e}}^j, e_i^j, u_{\mathrm{load}}^j, f_i^j, j, H)$
2:     $\Sigma_{s1} \leftarrow 0; \Sigma_{s2} \leftarrow 0$
3:     **while** $k = (j-1)T, (j-1)T+1, \ldots, jT,$ **do**
4:         Compute ζ_k by (11)
5:         $\Sigma_{s1} \leftarrow \Sigma_{s1} + \zeta_k \zeta_k^T; \Sigma_{s2} \leftarrow \Sigma_{s2}$
        $+ e(k-1)\zeta_k^T$
6:     **end while**
7:     $W^j = \frac{1}{T}\Sigma_{s1} - H; V_2^j = \frac{1}{T}\Sigma_{s2}$
8:     Obtain ξ_1^j and ξ_2^j via (17)
9:     **return** ξ_1^j, ξ_2^j
10: **end function**

Fig. 3. Four-area synthetic system with AGC in each area.

operation during some situations such as intentional tripping of load/generation [16]. Even without fine adjustments of frequency owing to AGC, the primary frequency control is capable of maintaining the system frequency within an acceptable range, say, from 59.96 Hz to 60.04 Hz [25], and the frequency falling into such a range will not trigger any load shedding events [26]. However, if stealthy cyberattacks on AGC are not detected in a timely fashion, they may keep compromising the control performance of the frequency regulation. For example, if a replay attack is not detected in time, the energy consumed by AGC actually bring no benefit to the grid in terms of regulating frequency, and the control performance of AGC is compromised.

IV. NUMERICAL EXAMPLES

This section presents the results on the efficacy of the dynamic-watermarking-based online defense algorithm on a four-area power system and the Northeastern Power Coordinating Council (NPCC) 140-bus power system. The malicious attacks to the synthetic system will be launched based on the attack templates presented in Section II-D. As will be shown, these attacks can be detected in a timely manner via the proposed approach without sacrificing the performance of the system.
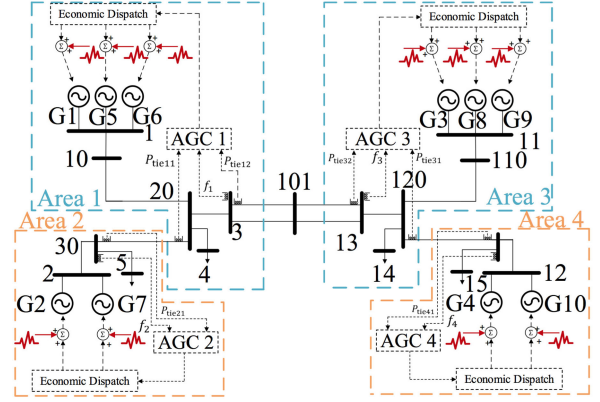
A. Performance Validation of the Proposed Algorithm on the Four-Area System

1) Four-Area System Description: This test system has four areas and ten generators, as shown in Fig. 3. The system is linearized about the given operating condition by Power System Toolbox (PST) [27], and the system matrices for the linear model, i.e., A, B and C in (1), are extracted. In order to mimic the behavior of AGC, in each area, we add a discrete proportional-integral (PI) feedback loop, where the proportional gain constant is set to -0.0745 and the integral gain is set to -0.0333. For each area, the PI controller takes its local measurements of tie-line power flows and frequency as its inputs and computes a control signal to change the load reference setpoint of the generator. This is done every 2 seconds, i.e., $\tau = 2$ and $\kappa_i \equiv 1$ for $i \in \{1, 2, 3, 4\}$. The load deviations around the scheduled values are modeled as independent and identically distributed (i.i.d.) Gaussian white noise with zero mean and covariance matrix $\sigma_L^2 I_8$, where I_8 is a 8×8 identity matrix. The variance $\sigma_L^2 = 0.0025$ is chosen such that the frequency fluctuates within the normal range, i.e., 60 ± 0.03 Hz [25] with high probability. The measurement noise of frequency and real power are normally distributed with zero mean. The variance of the frequency measurement noise, $\sigma_f^2 = 9.1891 \times 10^{-12}$, is tuned such that the accuracy of frequency measurement falls within ± 0.0005 Hz [28] with high probability, and the signal-to-noise ratio (SNR) of deviation measurements of tie-line flow is 20 dB. The covariance matrix of the process noise Q' is $10^{-9} I_{n'}$, where $I_{n'}$ is an identity matrix of dimension of n'.

2) Parameter Setting of the Proposed Algorithm: For the implementation of Algorithms 1 and 2, we have the following settings of the parameters:

- The number of samples in each block T is 30, so that ξ_1^j and ξ_2^j are computed every 60 seconds;
- The threshold η_1 is set to 2.5207×10^{-4} with $\xi_1^\infty = 3.6010 \times 10^{-5}$ and $\kappa' = 7$;
- the variance of the private injections σ_e in both Area 1 and Area 2 is set to 10^{-7}.

We first examine the impact of the private injection on the performance of the AGC in terms of frequency regulation. Fig. 4 records the control commands from AGC 1, and it shows that
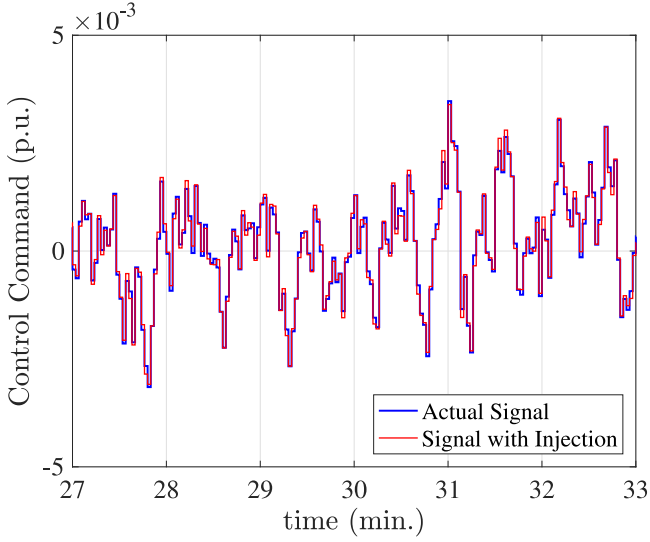
Fig. 4.   The impact of the private injection on the command signal showing that watermarking does not lead to any loss of performance under normal operation.
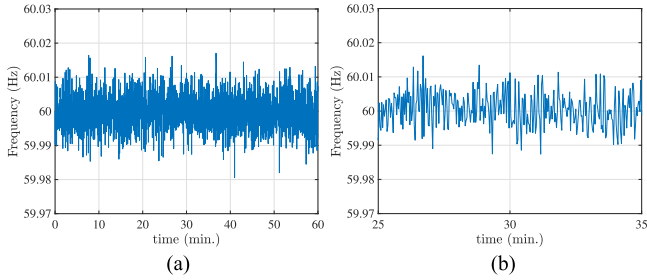

Fig. 5.   Frequency measurement (a) from 0 to 60 min, and (b) zoom-in frequency measurement from 25 to 35 min, under the **replay** attack to the frequency measurement of Area 1 launched at 30 min.

Fig. 6.   The evolutions of indicator ξ_1^j under the **replay** attack to the (a) **frequency** measurement and (b) **tie flow** measurement of **Area 1** starting at 30 min.

Fig. 7.   Frequency measurement in Area 1 (a) from 0 to 60 and (b) zoom-in frequency measurement from 25 to 35 min, under the **noise-injection** attack to the frequency measurement of Area 1 launched at 30 min.

Fig. 8.   The evolutions of indicator ξ_1^j under the **noise-injection** attack on (a) the **frequency** measurement, and (b) the **tie flow** measurement, of **Area 1** starting at 30 min.

the private injection does not cause significant deviation of the actual input from the control policy-specified input. The percentage of variance change of control command of AGC 1 and frequency are 0.26% and 1.73%, respectively, and the small change of the variance suggests negligible sacrifice of performance resulting from the private injection.

3) Detection of Replay Attack: We next demonstrate the efficacy of the dynamic watermarking approach for detecting replay attacks defined in Section II-D1. Fig. 5 shows the frequency measurements in Area 1. Beginning at 30 min, the frequency sensor reports a pre-recorded sequence of measurements instead of the actual measurements. No anomaly can be identified from Fig. 5, as no frequency constraint is violated within the time period of interest.

Next, the proposed Algorithms 1 and 2 are applied to detect the replay attack. In each area, the online detection algorithms compute the indicators ξ_1^j and ξ_2^j based on their local measurements of frequency and tie-line flow. The evolution of ξ_1^j over time in Area 1 is presented in Fig. 6(a). It is seen that ξ_1^j exceeds the threshold η_1 after 31 minutes, indicating that the attack starts between the 30th and 31st minutes. A similar result can be observed from Fig. 6(b) which presents the evolution of ξ_1^j under the replay attack to tie-line flow measurement of Area 1. After
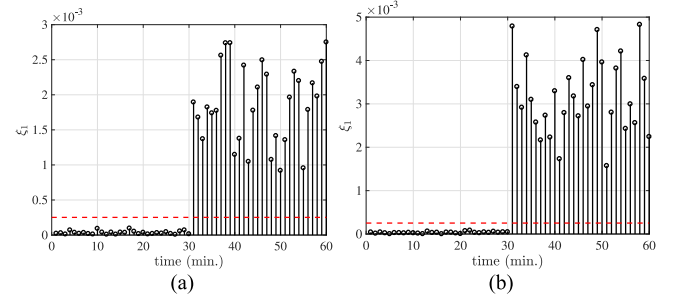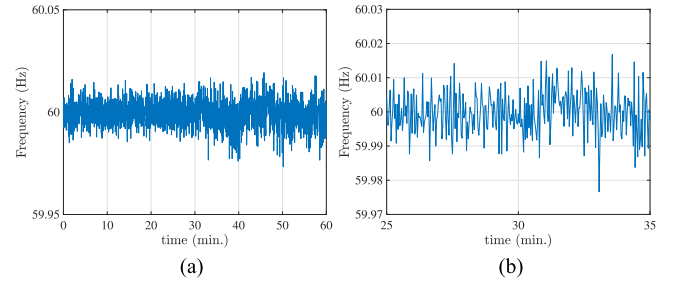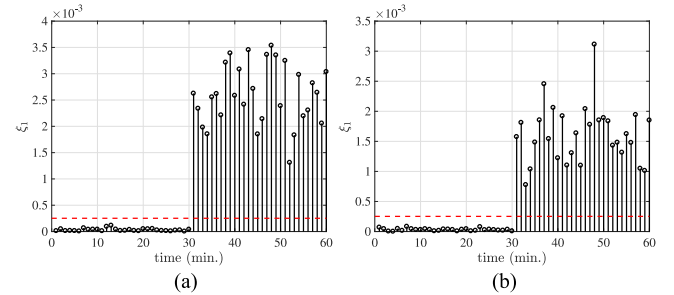
the attacks are detected, one mitigation action is to deactivate the AGC.

4) Detection of Noise-Injection Attack: In this section, we demonstrate the efficacy of the proposed approach for detection of noise-injection attacks. As mentioned in Section II-D, additional noise is superimposed on the actual frequency measurement after the 30th minute, and it is chosen so that the frequency is still within the normal range. Fig. 7 shows the measurements of the frequency before and after the attack, and, again, we cannot notice any anomaly since the frequency is within the normal range all the time and no distinct feature ever appears after 30 minutes. Using the proposed algorithm, the noise injection attack on the frequency measurements (Fig. 8) is identified successfully between the 30th and 31st minutes.

5) Detection of Destabilization Attack: This section deals with securing the system from destabilization attacks. A destabilization attack is carried out on the tie-line flow measurements
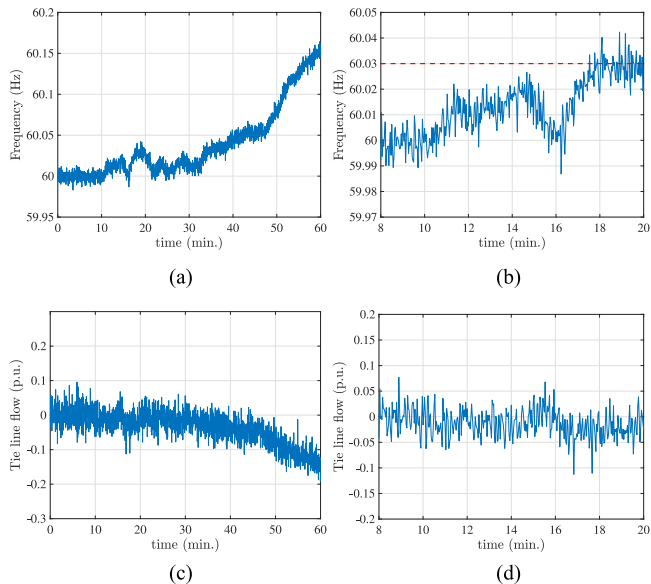
Fig. 9. Frequency measurement in (a) Area 1 from 0 to 60 min and (b) its zoom-in frequency measurement, (c) tie-line flow measurement in Area 1 from 0 to 60 min, and (d) its zoom-in tie-line flow measurement under the **destabilization** attack to the tie-line flow measurement of Area 1 launched at 10 min.

Fig. 10. The evolution of indicator ξ_1^j under the **destabilization** attack on the **tie flow** measurement of **Area 2** starting at 10 min.

Fig. 11. Frequency measurement under **destabilization attack** to (a) the **tie-line flow** measurement in Area 1, and (b) the evolution of corresponding ξ_1^j.

in Area 1. As mentioned in Section II-D3, the output sequence of a malicious filter M can be obtained via a simple tuning procedure, as follows. The adversaries may first force the sensor to report to the control center the scheduled tie-line flow plus a scaled version of actual flow deviation, i.e., $p_{\text{sch}} + \lambda \Delta p$ with an arbitrary chosen λ, as opposed to the actual flow measurement $p_{\text{sch}} + \Delta p$. Then the attackers can gradually tune λ such that the frequency exhibits unstable/oscillatory behavior. In the four-area system, the scalar λ is -0.89, and the attack starts at the 10th minute. Based on the scaled flow measurement Δp, the control command is computed according to the AGC control law, and the load reference setpoint of Generators 1, 5 and 6 are changed accordingly. As evident from Fig 9(a), the closed-loop system is unstable and the frequency grows in an unbounded fashion.

Now we observe the process of destabilization attack from the perspective of the system operator. Suppose that the system operator keeps monitoring the reported frequency and tie-line flow measurements at the balancing authority of Area 1. Then, Figs. 9(b) and (d) are what the operator can observe from the 8th minute to the 20th minute. The operator might not realize the anomaly until around the 16th minute at which time several samples of frequency exceed the upper limit of the normal frequency range. However, the proposed approach can detect the destabilization attack between the 10th minute and 11th minutes, as we can see from Fig. 10.

One might wonder if the ACE will always ultimately exceed its limits under a destabilization attack, in which case the operator will notice it anyway, thereby rendering the proposed approach superfluous. The answer is that there are sophisticated destabilization attacks where the ACE might not exhibit instability. Consider an attack template which is the same as
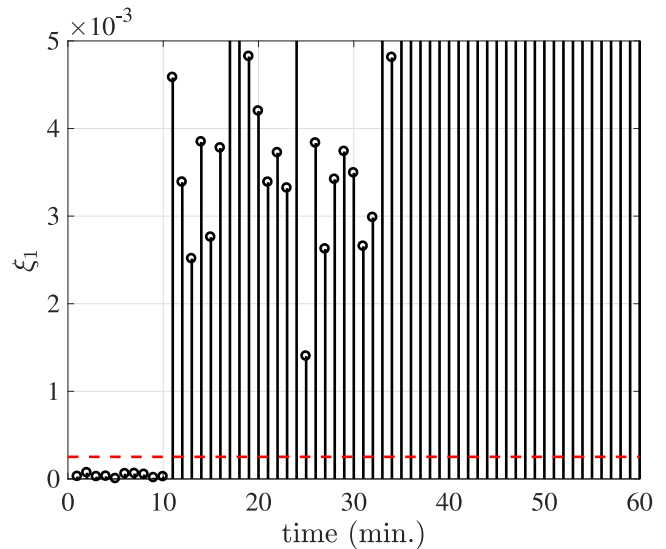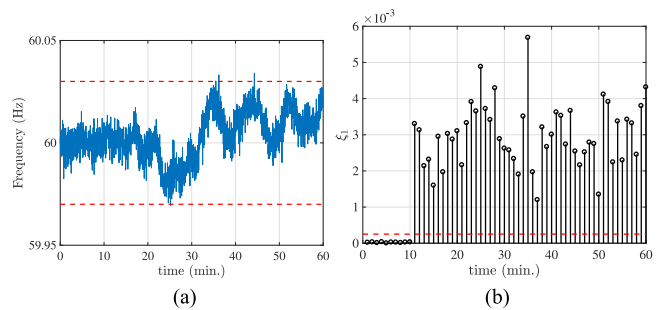
earlier, except that λ is set to -0.84. This results in the frequency measurement in Area 1 shown in Fig. 11(a). It can be seen that though some frequency samples exceed the constraint occasionally, these violations might be attributed to measurement error, and consequently be ignored by the operators since the frequency reverts to the normal range after several abnormal samples. In contrast the indicator signals under watermarking exhibit the consecutive spikes shown in Fig. 11(b) thereby detecting the attack on Area 1. It can be seen that, in contrast to performing fine adjustments of the system frequency, the energy consumed by AGC drives the frequency to oscillate within a wider range compared to the frequency before the cyber attack.

B. Performance Validation of the Proposed Algorithm on the NPCC 140-Bus System

1) NPCC 140-Bus System Description and Parameter Setting of the Proposed Algorithm: This benchmark system has 140 buses and 48 generators, and its raw parameters are available in the file named "`datanp48.m`" in PST [27]. In this paper, the NPCC 140-bus system is divided into two areas based on the geographical locations of buses [29], [30], and the bus indexes in each area are as reported in the Appendix. Accordingly,
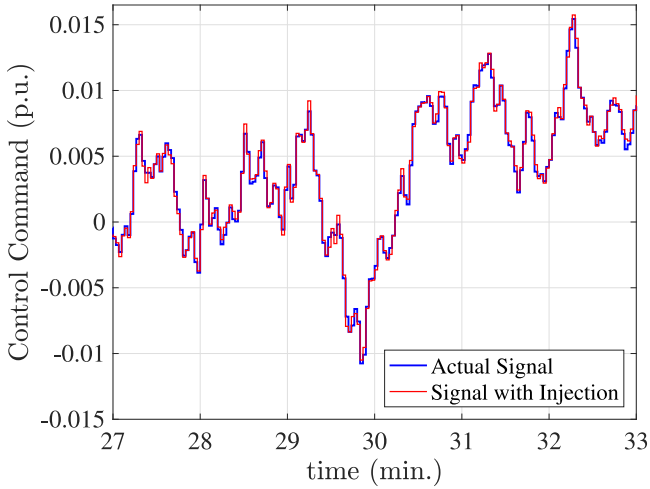
Fig. 12.   Control command comparison in the NPCC 140-bus power system.
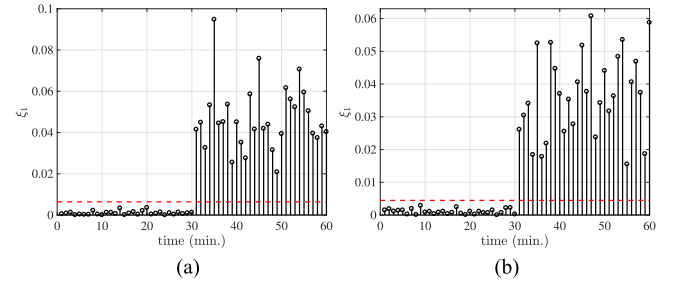

Fig. 13.   The evolutions of indicator ξ_1^j under (a) the replay attack and (b) the injection attack on the NPCC 140-bus power system.
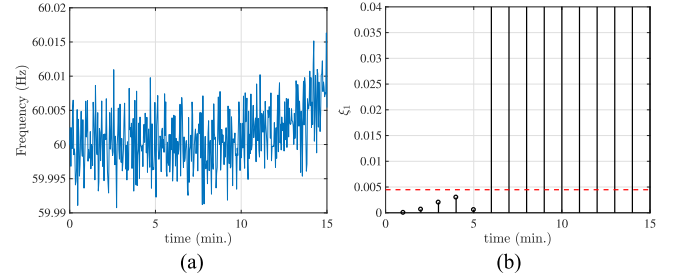

Fig. 14.   (a) Time-domain frequency measurements under the destabilization attack, (b) evolutions of indicator ξ_1^j under the destabilization attack.

eight transmission lines are chosen as the tie lines connecting the two areas; they are Line 78–81,[1] 76–77, 66–134, 67–138, 105–111, 105–106, 105–107, and 105–101. There are 9 generators in AGC loop, which are Generators 1, 2, 18, 19, 20, 21, 22, 23, and 24. The system matrices A, B, and C are extracted by PST. In Area 1, we add a discrete PI feedback loop, where both of the proportional gain constant and the integral gain are set to -0.0451. The variance parameter of the load deviations $\sigma_L^2 = 0.001$ is chosen such that the frequency fluctuates within the normal range, i.e., 60 ± 0.03 Hz [25] with high probability. The thresholds $\eta_1 = 0.0045$ with $\eta_1^\infty = 6.3935 \times 10^{-4}$ and $\kappa' = 7$. The settings of τ, κ_i, σ_f^2, Q', T, σ_e, and SNR of deviation measurements of tie-line flow are the same as those in Section IV-A.

Again, we examine the impact of the private injection on the performance of the AGC in terms of frequency regulation. Fig. 12 records the control commands from AGC. It shows that the control command with the private injection does not deviate significantly from the control policy-specified input.

2) Detection of Three Types of Cyber Attack: In this section, we demonstrate the efficacy of the proposed approach for detecting the three types of cyber attack defined in Section II-D, through simulations on the NPCC 140-bus power system. We first validate the performance of the proposed algorithm in terms of detecting the replay attack and the noise-injection attack on the frequency measurement in the NPCC 140-bus system. Both types of cyber attack begin at 30 min. As shown in Fig. 13, both types of cyber attack are identified successfully between the 30th and 31st minutes. We next deal with securing the NPCC 140-bus system from the destabilization attacks. The destabilization attack on the flow measurement of Line 78–81 starts at the 5th min, resulting in a growing trend of frequency deviation as shown in Fig. 14(a). The scalar λ defined in Section IV-A5 is -5. The evolution of ξ_1^j over time is presented in Fig. 14(b). It is observed that consecutive spikes exceed the threshold after the 6th min, suggesting that the attack appears between the 5th and 6th minutes.

---

[1]Line 78-81 represents the transmission line from Bus 78 to 81.

C. Comparison With the Regression-Based Approach

In this section, we compare the dynamic-watermarking approach with the regression-based approach [1] in the four-area system described in Section IV-A. In Reference [1], the cyber attacks on AGC are detected based on the following linear regression which characterizes the relationship between frequency (output) and load fluctuations (input), i.e.,

$$\hat{\omega}(k) \approx \sum_{h=0}^{H-1} \boldsymbol{\alpha}_h \boldsymbol{u}_{\text{load}}(k - h). \qquad (19)$$

Equation (19) assumes that the current frequency deviation $\hat{\omega}(k)$ is a linear combination of the current load fluctuation vector $\boldsymbol{u}_{\text{load}}(k)$ and the past load fluctuation vectors, i.e., $\boldsymbol{u}_{\text{load}}(k - h)$ for $h = 1, 2, \ldots, H-1$. $\boldsymbol{\alpha}_h$ is the combination coefficient vector, and the integer H is the order of the linear regression, which is an adjustable factor. The state-space version of (19) can be identified by the MATLAB System Identification Toolbox [31]. The attack is detected by checking the discrepancies between the reported frequency measurement $\omega(k)$ and its estimated value $\hat{\omega}(k)$. Hence, the indicator $\gamma(k)$ in the regression-based framework is defined by $\gamma(k) := \omega(k) - \hat{\omega}(k)$. An alarm is triggered if

$$|\gamma(k)| > \eta', \qquad (20)$$

where η' is the maximal $|\gamma(k)|$ under the normal condition or during the training stage [1].

However, the regression-based approach may not detect the following cyber attacks on AGC. The linear regression (19) can be learned by a sophisticated adversary, based on the input-output measurements. Then, the threshold η' can be approximately estimated. Finally, the actual measurement can be
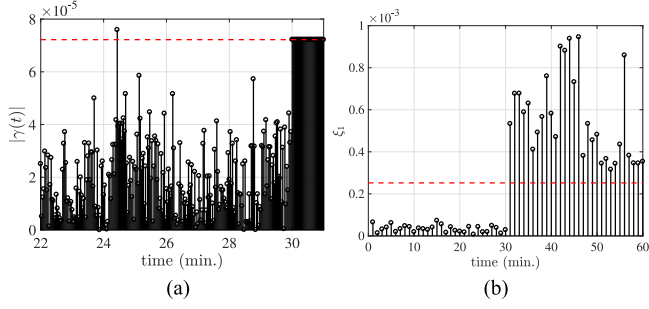
Fig. 15. The evolutions of (a) $|\gamma(t)|$ and (b) ξ_i^j over time under the attack defined in (21).

% of NRGs	NRG Index	θ_R	θ_I
0/9	N/A	6.5039	7.1692
1/9	24	6.4115	7.0572
2/9	23, 24	6.3088	6.9370
3/9	22, 23, 24	6.1079	6.6742
4/9	21, 22, 23, 24	6.1480	6.7020
5/9	20, 21, 22, 23, 24	6.0231	6.5594

replaced by the following malicious measurement sequence ω_a without being detected by the criteria (20):

$$\omega_{\mathrm{a}} = \hat{\omega} - \eta'. \qquad (21)$$

Next, we test the performance of the proposed algorithm in terms of detecting the attack defined in (21). The attack with $\eta' = 9.024 \times 10^{-5}$ starts at 30 min in the four-area system. Fig. 15(a) presents the evolution of the $|\gamma(t)|$ defined in the regression-based approach. It can be seen that $|\gamma(t)|$ does not exceed the threshold η' after the 30th minute, although it keeps being close to η'. In contrast, the indicator under the proposed method exceeds a predefined threshold consecutively after the 30th minute, suggesting that the attack defined in (21) can still be identified successfully, as shown in Fig. 15(b). Note that, although the regression-based approaches are not guaranteed to detect any cyber attacks, it can serve as a screening tool for the proposed framework.

D. Robustness Test

Due to the effect of deadband in generation units, some generators might not be responsive to small change in setpoints, and these generators are termed as non-responsive generators (NRGs). The number of non-responsive generators in AGC may impact the performance of the proposed framework. In order to investigate such an impact, we first define a performance indicator θ. In the context of Section IV-B, where the cyber attack (replay/injection) starts from the 30th min, the performance indicator θ can be defined as follows:

$$\theta = \frac{\min_p \xi_1^p}{\max_q \xi_1^q} \quad \forall p \in \{31, 32, \dots, 60\} \wedge q \in \{1, 2, \dots, 30\},$$
$$(22)$$

where the numerator suggests the minimal value of ξ_1^j under the attack, while the denominator is the maximal value of ξ_1^j under the normal condition. If the ratio $\theta > 1$, ξ_1^j under the attack can be linearly separated from that under the normal condition by setting a threshold, i.e, the attack can be detected by the proposed method.

In Section IV-B, we assume that all 9 generators in the AGC loop are responsive to small changes in their setpoints. Here, we increase the number of the NRGs from 0 to 5, and compute the corresponding performance indicators θ under the replay

attack and the noise-injection attack. The results are presented in Table I. It is seen that both θ_R and θ_I are greater than 1 under all scenarios, suggesting that the replay attack and the noise injection attack can still be detected, even though some non-responsive generators exist.

V. CONCLUSION

In this paper, an online framework to detect cyber attacks on AGC is proposed. In the proposed defense framework, a theoretically rigorous attack detection algorithm based on dynamic watermarking is developed to detect sophisticated adversaries equipped with extensive and even complete knowledge of the physical and statistical models of the power system. The proposed framework needs no hardware update of the generation units. The efficacy of the proposed framework is demonstrated in a four-area synthetic power system and a 140-bus power system. Future work will investigate the scaling up of the proposed method to larger-scale power systems.

APPENDIX A
DETERMINING THE THRESHOLD USING THE NEYMAN-PEARSON CRITERION

The detection thresholds can also be determined using the Neyman-Pearson criterion based on the maximum tolerable false alarm rate θ_0. Such a test is developed below.

We first note that the innovations process $\{\zeta\}$ in the paper is distributed according to $\zeta_k \sim \mathcal{N}(0, C_{di} P C_{di}^T + R)$ and i.i.d. across time (notation used is same as that in the paper). Therefore, it can be shown that $\zeta^T(k)\zeta(k) = \zeta_1^2(k) + \cdots + \zeta_n^2(k)$ is distributed according to the distribution of $\sum_{i=1}^{n} \lambda_i Z_i^2$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the covariance matrix $C_{di} P C_{di}^T + R$, and $Z_1, \dots, Z_n$ are i.i.d. Gaussian random variables with zero mean and unit variance. The distribution of $\zeta^T(k)\zeta(k)$, therefore, is that of a weighted linear combination of n independent chi-squared random variables with one degree of freedom each, and the distribution of the test statistic ξ_1^j in the paper is obtained by a $T-$ fold convolution of the above distribution (since it is the sum of T i.i.d. random variables) followed by a translation to make the resulting mean zero. Denote this distribution by f_0, and let H_0, the null hypothesis, be the case when the system is not under attack. Therefore,

$$H_0 : \xi_1^j \sim f_0,$$
$$H_1 : \xi_1^j \sim f_1,$$

where $f_1 \neq f_0$ is any arbitrary distribution.

Consider the threshold test

$$f_0(\xi_1^j) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_0$$

where γ_0 is the detection threshold, a value that is chosen in the Neyman-Pearson test depending on θ_0, the maximum tolerable false alarm rate. The false alarm rate for the test is given by

$$\mathbb{P}(H_1|H_0) = \int_{x:f_0(x)<\gamma} f_0(x)dx,$$

which when viewed as a function of γ_0, is monotonically increasing. The optimal threshold γ_0^* is determined as

$$\arg\max \int_{x:f_0(x)<\gamma_0} f_0(x)dx$$

$$\text{s.t.} \quad \int_{x:f_0(x)<\gamma_0} f_0(x)dx < \theta_0.$$

Since the function f_0 is known, the threshold η_1 in the paper can be determined from γ_0^*. A similar procedure is followed to determine the value of the threshold η_2.

There are two notions of detection performance in this context: (i) detection delay, and (ii) false alarm rate. These are conflicting objectives, and the problem of designing "optimal" detection schemes in our context is a special instance of this more general "quickest change detection" problem, and in what follows, we outline how exactly our problem maps to the latter. Further analysis of detection performance, such as optimal trade-offs between the detection delay and the false alarm rate, or optimal detection mechanisms for various objective functions, are well-studied problems in the quickest change detection literature (see [32]–[36] and references therein).

The quickest change detection problem has three ingredients [36]: (i) a stochastic process that is under observation, (ii) a random time τ at which the statistics of the process changes from one distribution to another, and (iii) a detection algorithm that observes the stochastic process and declares at each time t whether $t \geq \tau$ (change has occurred) or $t < \tau$ (change has not yet occurred). This problem has a long history with applications in various fields such as manufacturing systems, quality control, network security, econometrics, etc.

In the context of our paper, the innovations process corresponding to the state estimate process $\{\hat{x}_{di}(k|k)\}$ is the stochastic process that is under observation or test, and the time at which the statistics of the process changes is the time at which the attack is initiated. As prior works show [13], [14], in the presence of dynamic watermarking, the adversary cannot introduce any significant distortion without causing a change in the statistics of the stochastic process under test, no matter what attack strategy it chooses to employ. Thus, formulating our problem as one of quickest change detection provides us with a mature framework for developing different statistical tests that are optimal for various objective functions.

APPENDIX B
PARAMETERS OF THE FOUR-AREA POWER SYSTEM AND BUS INFORMATION OF EACH AREA IN THE NPCC 140-BUS SYSTEM

The four-area system used in Section IV-A is modified from the benchmark two-area system [16], [27]. The information on the parameters of buses and transmission lines of the four-area system is reported in [37]. In the four-area system, all generators associated with their exciters, governors and power system stabilizers (PSS) are the same, and their parameters can be found in "d2asbegp.m" in PST [27].

The NPCC 140 bus system is divided into two areas in this paper. The buses in the Area 1 are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 48, 49, 50, 51, 52, 53, 54, 56, 57, 58, 59, 60, 61, 62, 63, 64, 64, 65, 66, 67, 68, 70, 71, 72, 73, 74, 77, 78, 79, 80, 91, 92, 93, 94, 95, 96, 97, 99, 100, 101, 102, 103, 104, 106, 107, 108, 109, 110, 111, and 140. The rest of buses are in the Area 2.

REFERENCES

[1] R. Tan *et al.*, "Modeling and mitigating impact of false data injection attacks on automatic generation control," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 7, pp. 1609–1624, Jul. 2017.

[2] S. Sridhar and M. Govindarasu, "Model-based attack detection and mitigation for automatic generation control," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 580–591, Mar. 2014.

[3] A. Ashok, P. Wang, M. Brown, and M. Govindarasu, "Experimental evaluation of cyber attacks on automatic generation control using a CPS security testbed," in *Proc. IEEE Power Energy Soc. Gen. Meet.*, Jul. 2015, pp. 1–5.

[4] A. Ashok, S. Sridhar, A. D. McKinnon, P. Wang, and M. Govindarasu, "Testbed-based performance evaluation of attack resilient control for AGC," in *Proc. Resilience Week*, Aug. 2016, pp. 125–129.

[5] C. L. DeMarco, "Design of predatory generation control in electric power systems," in *Proc. 31st Hawaii Int. Conf. Syst. Sci.*, 1998, vol. 3, pp. 32–38.

[6] H. E. Brown and C. L. DeMarco, "Risk of cyber-physical attack via load with emulated inertia control," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2017.2697823.

[7] "Dragonfly: Cyberespionage attacks against energy suppliers," Symantec Security Response White Paper, Jul. 17, 2014. [Online]. Available: https://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/Dragonfly_Threat_Against_Western_Energy_Suppliers.pdf

[8] "Stuxnet." [Online]. Available: https://en.wikipedia.org/wiki/Stuxnet. Accessed on: May 15, 2018.

[9] M. Q. Ali, R. Yousefian, E. Al-Shaer, S. Kamalasadan, and Q. Zhu, "Two-tier data-driven intrusion detection for automatic generation control in smart grid," in *Proc. IEEE Conf. Commun. Network Security*, Oct. 2014, pp. 292–300.

[10] A. Ashok, M. Govindarasu, and J. Wang, "Cyber-physical attack-resilient wide-area monitoring, protection, and control for the power grid," *Proc. IEEE*, vol. 105, no. 7, pp. 1389–1407, Jul. 2017.

[11] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM Conference on Computer and Communications Security* (CCS '09). New York, NY, USA: ACM, 2009, pp. 21–32, doi: 10.1145/1653662.1653666.

[12] Z. Guo, K. H. Johansson, and L. Shi, "A study of packet-reordering integrity attack on remote state estimation," in *Proc. 35th Chinese Control Conf.*, Jul. 2016, pp. 7250–7255.

[13] B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber physical systems," *Proc. IEEE*, vol. 105, no. 2, pp. 219–240, Feb. 2017.

[14] B. Satchidanandan and P. R. Kumar, "On minimal tests of sensor veracity for dynamic watermarking-based defense of cyber-physical systems," in *Proc. 9th Int. Conf. Commun. Syst. Netw.*, Jan. 2017, pp. 23–30.

[15] T. Huang, M. Wu, and L. Xie, "Prioritization of PMU location and signal selection for monitoring critical power system oscillations," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 3919–3929, Jul. 2018.

[16] P. Kundur, N. J. Balu, and M. G. Lauby, *Power System Stability and Control*. New York, NY, USA: McGraw-Hill, 1994, vol. 7.

[17] Ibraheem, P. Kumar, and D. P. Kothari, "Recent philosophies of automatic generation control strategies in power systems," *IEEE Trans. Power Systems*, vol. 20, no. 1, pp. 346–357, Feb. 2005.

[18] J. Carpentier, "To be or not to be modern' that is the question for automatic generation control (point of view of a utility engineer)," *Int. J. Elect. Power Energy Syst.*, vol. 7, no. 2, pp. 81–91, 1985.

[19] A. A. Thatte, F. Zhang, and L. Xie, "Frequency aware economic dispatch," in *Proc. North Amer. Power Symp.*, Aug. 2011, pp. 1–7.

[20] J. S. Bay, *Fundamentals of Linear State Space Systems*. New York, NY, USA: McGraw-Hill Science, Engineering & Mathematics, 1999.

[21] *Matlab Documentation*, The MathWorks Inc., 2017. [Online]. Available: https://www.mathworks.com/help/control/ref/feedback.html. Accessed on: May 15, 2018.

[22] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2009, pp. 911–918.

[23] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 4, pp. 1396–1407, Jul. 2014.

[24] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Philadelphia, PA, USA: SIAM, 2015.

[25] EPRI, *EPRI Power System Dynamic Tutorial*. Palo Alto, CA, USA: Electric Power Research Institution, Jul. 2009.

[26] "Standard prc-006-npcc-1 automatic underfrequency load shedding," North American Electric Reliability Corporation, Atlanta, GA, USA, Tech. Rep. [Online]. Available: http://www.nerc.com/files/PRC-006-NPCC-1.pdf

[27] J. H. Chow and K. W. Cheung, "A toolbox for power system dynamics and control engineering education and research," *IEEE Trans. Power Syst.*, vol. 7, no. 4, pp. 1559–1564, Nov. 1992.

[28] Y. Zhang *et al.*, "Wide-area frequency monitoring network (FNET) architecture and applications," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 159–167, Sep. 2010.

[29] J. Qi, K. Sun, and W. Kang, "Optimal PMU placement for power system dynamic state estimation by using empirical observability gramian," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 2041–2054, Jul. 2015.

[30] C. Liu, B. Wang, F. Hu, K. Sun, and C. L. Bak, "Online voltage stability assessment for load areas based on the holomorphic embedding method," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 3720–3734, Jul. 2018.

[31] L. Ljung, "System identification toolbox," in *The Matlab User's Guide*. Natik, MA, USA: MathWorks 1988.

[32] A. G. Tartakovsky and V. V. Veeravalli, "General asymptotic Bayesian theory of quickest change detection," *Theory Probab. Its Appl.*, vol. 49, no. 3, pp. 458–497, 2005.

[33] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, Dec. 1971, doi: 10.1214/aoms/1177693055.

[34] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, no. 1, pp. 206–227, 03 1985, doi: 10.1214/aos/1176346587.

[35] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory Probab. Appl.*, vol. 8, no. 1, pp. 22–46, 1963, doi: 10.1137/1108002.

[36] V. V. Veeravalli and T. Banerjee, "Chapter 6—Quickest change detection," in *Academic Press Library in Signal Processing: Volume 3* (Academic Press Library in Signal Processing), A. M. Zoubir, M. Viberg, R. Chellappa, and S. Theodoridis, Eds. Amsterdam, The Netherlands: Elsevier, 2014, vol. 3, pp. 209–255. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124115972000060

[37] T. Huang, "Bus and branch information of a modified four-area power system," 2018. [Online]. Available: http://dropbox.com/s/j84badyisfntz4u/Four-area%20system.pdf?dl=0. Accessed on: May 15, 2018.
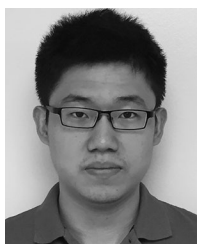
Tong Huang (S'16) received the B.E. degree in electric power engineering and its automation from North China Electric Power University, Baoding, China, in 2013 and the M.S. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2017, where he is currently working toward the Ph.D. degree. His research interests include power system stability, security, and wide-area monitoring and control systems.

Bharadwaj Satchidanandan received the Master's degree from the Indian Institute of Technology Madras, Chennai, India, where he worked on wireless communications. He is working toward the Doctoral degree in the Electrical and Computer Engineering Department, Texas A&M University, College Station, TX, USA. Between May 15 and August 15, he interned at Intel Labs, Santa Clara, CA, USA, where he worked on interference cancellation algorithms for next-generation wireless networks. His research interests include cyberphysical systems, power systems, security, database privacy, communications, control, and signal processing.

P. R. Kumar (LF'18) received the B.Tech. degree from IIT Madras, Chennai, India, in 1973, the D.Sc. degree from Washington University in St. Louis, St. Louis, MO, USA, in 1977. He was a faculty member with University of Maryland, Baltimore County (1977–1984) and University of Illinois, Urbana-Champaign (1985–2011). He is currently with Texas A&M University, College Station, TX, USA. His current research is focused on stochastic systems, energy systems, wireless networks, security, automated transportation, and cyberphysical systems. He is a member of the U.S. National Academy of Engineering and The World Academy of Sciences. He was awarded a Doctor Honoris Causa by ETH, Zurich. He was the recipient of the Award for Control Systems, the Donald P. Eckman award, the IEEE Field Award of the AACC, the Fred W. Ellersick Prize of the IEEE Communications Society, the Outstanding Contribution Award of ACM SIGMOBILE, the INFOCOM Achievement Award, the SIGMOBILE Test-of-Time Paper Award, the Distinguished Alumnus Award from IIT Madras, the Alumni Achievement Award from Washington University, and the Daniel Drucker Eminent Faculty Award from the College of Engineering at the University of Illinois. He is an ACM Fellow. He was Leader of the Guest Chair Professor Group on Wireless Communication and Networking at Tsinghua University, is a D. J. Gandhi Distinguished Visiting Professor with IIT Bombay, and an Honorary Professor with IIT Hyderabad.
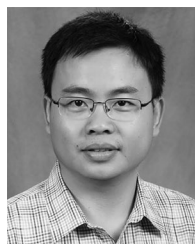
Le Xie (S'05–M'10–SM'16) received the B.E. degree in electrical engineering from Tsinghua University, Beijing, China, in 2004, the M.S. degree in engineering sciences from Harvard University, Cambridge, MA, USA, in 2005, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, in 2009. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. His research interests include modeling and control of large-scale complex systems, smart grids application with renewable energy resources, and electricity markets.