Dynamic Anomaly Detection with High-fidelity Simulators
A Convex Optimization Approach

Pan, Kaikai; Palensky, Peter; Esfahani, Peyman Mohajerin

DOI
10.1109/TSG.2021.3129074

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Smart Grid

Dynamic Anomaly Detection With High-Fidelity Simulators: A Convex Optimization Approach

Kaikai Pan, *Member, IEEE*, Peter Palensky, *Senior Member, IEEE*, and Peyman Mohajerin Esfahani

***Abstract*—The main objective of this article is to develop scalable dynamic anomaly detectors with high-fidelity simulators of power systems. On the one hand, models in high-fidelity simulators are typically "intractable" if one opts to describe them in a mathematical formulation in order to apply existing model-based approaches from the anomaly detection literature. On the other hand, pure data-driven methods developed primarily in the machine learning literature neglect our knowledge about the underlying dynamics of power systems. In this study, we combine tools from these two mainstream approaches to develop a data-assisted model-based diagnosis filter utilizing both the knowledge from a picked abstract model and also the data of simulation results from high-fidelity simulators. The proposed diagnosis filter aims to achieve two desired features: (i) performance robustness with respect to model mismatch; (ii) high scalability. To this end, we propose a tractable (convex) optimization-based reformulation in which decisions are the filter parameters, the model-based information introduces feasible sets, and the data from the simulator forms the objective function to-be-minimized regarding the effect of model mismatch on the filter performance. To validate the theoretical results, we implement the developed diagnosis filter in DIgSILENT PowerFactory to detect false data injection attacks on the Automatic Generation Control measurements in the three-area IEEE 39-bus system.**

***Index Terms*—Anomaly detection, data-assisted model-based, diagnosis filter, model mismatch, high-fidelity simulator, convex optimization.**

I. Introduction

THE PRINCIPLE of anomaly detection in power system cyber security is to generate a diagnostic signal (e.g., residual) that keeps sensitive to malicious intrusions and simultaneously robust against other unknowns, given the available data from system outputs [1], [2]. The detection methods can be mainly classified into two categories: (i) the approaches that exploit an explicit mathematical model of the system dynamics (referred to *model-based* methods in this article); (ii) the *data-driven* approaches that try to automatically learn the system characteristics from the output data [3], [4]. Our work in [5] has developed a scalable diagnosis tool to detect the class of multivariate false data injection (FDI) attacks which may remain stealthy in view of a static detector, by capturing the dynamics signatures of such a disruptive intrusion. This method is, indeed, model-based that the dynamics of the system trajectories under multivariate FDI attacks are described via an explicit mathematical model representation (i.e., linear differential-algebraic equations (DAEs)). The numerical results in [5] have proven its effectiveness in the given linear mathematical model. Now here comes another question:

Can the power of scalable model-based diagnosis tools be still utilized in real-world applications such as electric power systems for which there are reliable datasets from high-fidelity but complex simulators?

We aim to address the question by designing a scalable diagnosis tool with high-fidelity simulators. To do that, we need two important pieces of information: (i) the knowledge of an abstract model; (ii) the system trajectories provided by a high-fidelity simulator. The abstract model-based knowledge is utilized for a scalable design such that the design parameters (e.g., the order of the diagnosis tool) should be adjustable depending on the size or degree of the studied system. The high-fidelity simulator is a stand-in for actual measurements to be fitted into the diagnosis tool. Let us further clarify the terminologies adopted above. An abstract model refers to an explicit, but perhaps reduced-order, mathematical description of power system dynamics, e.g., a linear DAE. The actual measurements now refer to the reliable datasets of simulation results from a high-fidelity simulator like DIgSILENT PowerFactory. A simulator is said to be high-fidelity when it is the closest to the reality and may consist of several complex nonlinear DAEs as parts [6]. However, unfortunately, one may not have access to the mathematical description of such simulator. The reasons behind come from many aspects, but particularly, many high-fidelity simulators are still commercial. Then, it is not hard to observe that the source of challenge to answer the question above comes from the following aspect: whatever abstract model we can pick, *model mismatch* is always reflected through the difference of the output of the abstract model and the one from the high-fidelity simulator. It can be expected that this unknown of model mismatch potentially affects the diagnostic performance. With that in mind, we propose a scalable diagnosis tool that is robust with respect to model mismatch, by exploiting the information revealed to us

Manuscript received May 5, 2021; revised September 4, 2021; accepted November 7, 2021. Date of publication November 18, 2021; date of current version February 21, 2022. The work of Kaikai Pan was supported by the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform) under Project ZJUNGICS2021024. The work of Peyman Mohajerin Esfahani was supported by the ERC under Grant TRUST-949796. Paper no. TSG-00710-2021. *(Corresponding author: Kaikai Pan.)*

Kaikai Pan is with the College of Electrical Engineering, Zhejiang University, Hangzhou 310058, China (e-mail: pankaikai@zju.edu.cn).

Peter Palensky is with the Department of Electrical Sustainable Energy, Delft University of Technology, 2600 AA Delft, The Netherlands (e-mail: p.palensky@tudelft.nl).

Peyman Mohajerin Esfahani is with the Delft Center for Systems and Control, Delft University of Technology, 2600 AA Delft, The Netherlands (e-mail: p.mohajerinesfahani@tudelft.nl).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2021.3129074.

Digital Object Identifier 10.1109/TSG.2021.3129074

through the simulation data together with the abstract-model based knowledge, resulting a novel data-assisted model-based design perspective. We will provide further details toward this objective in Section II.

Literature on model-based and data-driven anomaly detection: Let us briefly overview the advantages and limitations of the pure model-based and data-driven approaches. The model-based methods require detailed information of the studied system. Some research papers deploy the statistical properties of system outputs, such as the work in [7] using cumulative sum-type algorithms for a sequential detection and the one in [8] using the measurement consistency assessment. These techniques can be essentially confined by some prior assumptions on system output errors. Recently, approaches based on moving target defense have been proposed to actively change the system configuration to detect various cyber attacks [9], [10], while the additional defense cost has to be considered. Another major subclass of these schemes is the observer-based residual generator that historically emerges from a control-theoretic perspective and has been extended to linear DAEs by [11]. To our best of knowledge, the study [12] is the first attempt to apply observer-based detectors to power system cyber security problem. Recently, a variant of observer-based method is employed in [13] so as to deal with unknown exogenous inputs in the linear Automatic Generation Control (AGC) system. Parameter estimation model-based approaches have also been extensively investigated. For instance, the extended Kalman filter algorithm is used to perform such an estimation for anomaly detection [14]. In [15], a comparison study is carried out for various Kalman filters and observers in power system dynamic state estimation with model uncertainties and malicious cyber attacks. The residual generators above usually have the same degree as the system dynamics, which can be problematic in the online implementation particularly for large-scale power systems [16]. Our diagnosis filter in [5] provides a good alternative to detect multivariate FDI attacks in a real-time operation. Still, the challenge remains as the power system models are mostly nonlinear, complex and high-dimensional. The work in [17] proposed an optimization-based filter for detecting a single anomaly in the control system where the nonlinearity can be fully described in DAEs. However, as noted earlier, having a detailed mathematical description of the model especially in the high-fidelity simulator or a real electric power system is usually infeasible.

Another major technique for anomaly detection comes from data-driven approaches that do not require an explicit mathematical model of system dynamics. Developments such as sensing technology, Internet-of-Things and Artificial Intelligence have contributed to a more data-driven power system [18]. Anomaly detection is mainly considered as a classification problem and there are supervised, unsupervised or semi-supervised learning methods for that purpose. Many efforts have been made on supervised classifications among which deep neural networks (DNN) [19], [20], bayesian networks [21] and support vector machines (SVM) [22] are the popular approaches. For unsupervised classifications to detect cyber attacks in smart grids, one can find principle component analysis (PCA) and its extension [23], autoencoders [24], etc. Some other research works have developed semi-supervised learning type anomaly detectors: in [25], a semi-supervised SVM is first proposed; a semi-supervised mixture Gaussian distribution based formulation is introduced in [26]; the recent study [27] shows a promising semi-supervised method by integrating the autoencoders into an advanced generative adversarial network (GAN). In addition to the approaches above, it is noteworthy that the study of [28] has deployed a reinforcement learning based algorithm for online attack detection in smart grids without a prior knowledge of system models or attack types. Overall, data-driven methods are suitable for real implementations in complex and large-scale systems. However, their performance highly depends on the quantity and quality of the accessible data[1], and thus can be intractable in many cases [29]. Besides, the required pre-processing stage (e.g., data training) may have a high computational cost.

Contributions and outline: This article aims to develop a scalable and robust diagnosis filter with high-fidelity simulators like PowerFactory. To achieve that, we propose a tractable optimization-based reformulation where the abstract model-based information introduces feasible sets, and the simulation data forms the objective function to minimize the effect of model mismatch on the filter residual. In this way, the diagnosis filter can be "trained" in the normal operations (without attacks) to have performance robustness with respect to model mismatch. Then it can be "tested" in PowerFactory to detect FDI attacks. Our main contributions are:

(i) Firstly, we develop a data-assisted model-based approach that utilizes both the model-based knowledge and also the simulation data from the simulator, for a scalable and robust design (Definition 2 and the program (9)). Instead of using any existing machine learning algorithms, we propose our own optimization-based characterization to "train" the filter under multiple mismatch signatures obtained through the simulation data (Remark 2). As far as we know, this is the first study that builds on such a perspective. In the optimization-based reformulation, the objective is to minimize the effect of model mismatch on the filter residual. We show that, the resulted optimization programs are convex and hence tractable, indicating that the proposed filter is not computational expensive compared to many pure data-driven methods.

(ii) We investigate optimization-based characterizations of the developed diagnosis filter in both scenarios of univariate and multivariate attacks. A square of $\mathcal{L}_2$-inner product with corresponding norm is proposed to quantify the effect of model mismatch. Then the $\mathcal{L}_2$-norm of the residual part introduced by model mismatch is reformulated as a quadratic function. We prove that, the characterization of the filter under a univariate attack becomes a family of convex quadratic programs (QPs), and the developed filter can even have the capability of tracking the attack magnitude through its non-zero steady-state

[1] To be noted, there are methods like the semi-supervised ones that can help in reducing the quantity of labeled data for the training [27].

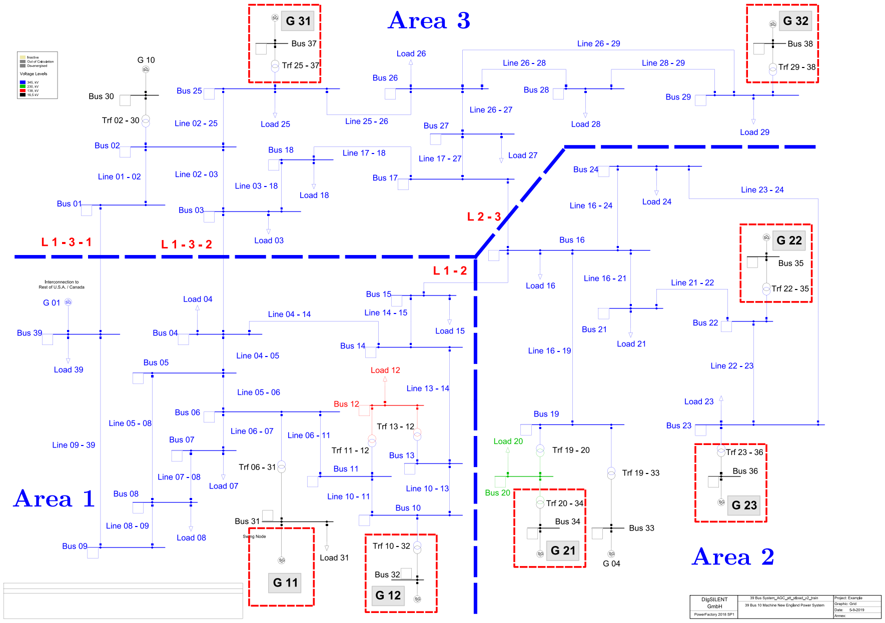


Fig. 1. Three-area power system with AGC functions modeled in DIgSILENT PowerFactory.

residual (Theorem 1). We also extend to the scenario of multivariate attack, and it appears that a standard QP can be derived. Besides, we provide conditions under which the filter can detect all the plausible multivariate attacks in an admissible set (Corollary 1).

The process of diagnosis filter construction and validation is concluded in Algorithms 1 and 2. The effectiveness of the proposed approach is validated on the three-area IEEE 39-bus system. Numerical results from the case study illustrate that the developed data-assisted model-based filter with the PowerFactory simulator can successfully generate alerts in the presence of disruptive univariate or multivariate FDI attacks, while a pure model-based detector may fail.

Section II shows the outline of our proposed solution. Both the motivating case study and the mathematical framework of our solution are detailed presented. Section III proposes the tractable optimization-based characterization of the developed diagnosis filter which has high scalability and performance robustness to model mismatch. Numerical results of the developed filter comparing with other pure model-based ones are reported in Section IV. Conclusions are drawn in Section V.

II. Outline of the Proposed Solution

In this section, we start from a case study motivating the setting of our work. Then the mathematical framework of our proposed solution to design a diagnosis filter that can be applied to the high-fidelity simulator is presented.

A. Motivating Case Study

For a detailed description of a real electric power system, high-fidelity simulators are always exploited to predict its behavior. Let us consider a multi-area power system equipped with AGC functions in Figure 1. The three-area IEEE 39-bus system is modeled in DIgSILENT PowerFactory. In PowerFactory, the dynamic generator model would consist of a synchronous machine, along with a voltage regulator for the exciter, and also the turbine-governor unit. We also implement AGC in each area for secondary frequency control. The AGC is a typical automatic control loop that regulates the system frequency and the power exchanges between areas by controlling the power settings of the generators participating in AGC to follow load changes. Looking into the system behavior of electromechanical dynamics, we can describe the formal mathematical model of the studied system in the high-fidelity simulator by a set of DAEs,

$$\dot{\boldsymbol{x}}(t) = \mathsf{f}(\boldsymbol{x}(t), \boldsymbol{a}(t), \boldsymbol{f}(t)) \tag{1a}$$

$$\boldsymbol{0} = \mathsf{g}(\boldsymbol{x}(t)\,\boldsymbol{a}(t)\,\boldsymbol{d}(t)), \tag{1b}$$

where $\boldsymbol{x} \in \mathbb{R}^{n_x}$ is the vector of augmented state variables including the ones of synchronous machines (e.g., damper-windings, mechanical equations of motion, exciter, voltage regulator, turbine-governor unit) and AGC controllers. The vector $\boldsymbol{a} \in \mathbb{R}^{n_a}$ represents the algebraic variables. The vector $\boldsymbol{d} \in \mathbb{R}^{n_d}$ denotes the (load) disturbances. The vector $\boldsymbol{f} \in \mathbb{R}^{n_f}$ characterizes the possible anomalies that affect the dynamics of system trajectories, which will be detailed in Section II-B. We note that (1a) consists of both synchronous generator dynamics and AGC controller dynamics of the multi-area power system in PowerFactory.

In the nonlinear DAE formulation of (1), the function $\mathsf{f} : \mathbb{R}^{n_x+n_a+n_f} \rightarrow \mathbb{R}^{n_x}$ in (1a) captures the dynamics of synchronous generators and AGC controllers. The function $\mathsf{g} : \mathbb{R}^{n_x+n_a+n_d} \rightarrow \mathbb{R}^{n_a}$ in (1b) models the electrical network. These nonlinear functions involve saturation, sinusoidal terms and possibly other nonlinearity in the power system. Since (1) is for a model description in the high-fidelity simulator like PowerFactory, one may not have access to the detailed DAE of (1). In what follows, we show that, to design a diagnosis filter with the high-fidelity simulator, we do not need an explicit model of (1), however, it is this simulator that gives us the trajectory of the system and the output from the simulations to be fitted into the diagnosis filter for anomaly detection.

First, if the model in (1) is known, a common practice for the analysis is to linearize (1) under an assumption that the system works closely around the nominal operating point; one can find such treatments in many literature sources like [12], [30]. Next, though (1) in PowerFactory is unknown to the diagnosis filter design, we can pick an model whose parameters are obtained from a simplified version of the system. If the mathematical description of the complex DAE model (1) would be available, then a reasonable choice of this abstract model could be the linearized DAE around the operating point. However, we emphasize that here *this choice does not need to be a linearized version of the DAE*, particularly in view of the robustification technique introduced in the next step. Such an abstract model can be picked as

$$\begin{aligned}\dot{\tilde{\boldsymbol{x}}}(t) &= \boldsymbol{A}_{c,x}\tilde{\boldsymbol{x}}(t) + \boldsymbol{B}_{c,d}\boldsymbol{d}(t) + \boldsymbol{B}_{c,f}\boldsymbol{f}(t),\\ \boldsymbol{y}(t) &= \boldsymbol{C}\tilde{\boldsymbol{x}}(t) + \boldsymbol{D}_f\boldsymbol{f}(t),\end{aligned} \tag{2}$$

where $\tilde{\boldsymbol{x}} \in \mathbb{R}^{n_{\tilde{x}}}$ is the state vector of the abstract model including the dynamics of the multi-area system and also AGC. The vector $\boldsymbol{y} \in \mathbb{R}^{n_y}$ denotes the system output of (2). In (2), $\boldsymbol{A}_{c,x} \in \mathbb{R}^{n_{\tilde{x}}\times n_{\tilde{x}}}$ is the state matrix; $\boldsymbol{B}_{c,d} \in \mathbb{R}^{n_{\tilde{x}}\times n_d}$ and $\boldsymbol{B}_{c,f} \in \mathbb{R}^{n_{\tilde{x}}\times n_f}$ relate disturbances and anomalies to the system; $\boldsymbol{C} \in \mathbb{R}^{n_y\times n_{\tilde{x}}}$ is the output matrix; $\boldsymbol{D}_f \in \mathbb{R}^{n_y\times n_f}$ characterizes the corrupted system outputs by anomalies.

The model (2) is picked because, as noted by [31], in AGC we pay more attention on collective performance of all generators, and hence we assume that generators in each area have the same characteristics and each area can be represented by an aggregated model comprised of equivalent turbine-governor units and generators [32]. AGC acts as the secondary frequency control of a multi-area system, and has relatively slow dynamics. Thus, from the timescales of interest, the frequency response of AGC can be decoupled from the loop of automatic voltage regulator. This is due to the fact that the time constant of voltage regulator dynamics is quite smaller than that of AGC [13]. It is feasible to use a quasi-state model that assumes a steady-state operating point of the voltage regulator loop ignoring its fast dynamics. Correspondingly, $\tilde{x}$ in (2) may be a reduced-order one comparing with x in (1). Of course, the model (2) is not meant to be "low-fidelity", instead, as explained above regarding the interested collective performance of generators and the time scales in system dynamics, the model (2) decoupled from voltage regulator loops with certain level of abstractions can be sufficiently accurate especially for analytical analysis [5], [13]. However, as mentioned in Section I, the challenge is, regardless of the choice of the abstract model (2), there would always exist *model mismatch* between (1) and (2).

Remark 1 (Model Mismatch Sources): The sources of model mismatch between (1) and (2) emerge from various aspects. For instance, we notice that the assumption of operating point can be violated due to switches and disturbances. Some nonlinear parts like saturation, sinusoidal terms are not considered in (2), resulting another part of model reduction "error". Various types of inputs, such as processing and measurement noises, parameter variations, bad data, along with the exogenous ones of load changes and cyber attacks particularly noted in (1) and (2), could also contribute to the model mismatch.

B. Anomalies: Univariate or Multivariate FDI Attacks

We continue to present the anomaly scenario of our motivating case study. In the AGC of a multi-area system, the AGC controller collects the information of grid frequency and power exchange on the tie-line (e.g., L1-2 in Figure 1) that connects areas to form the area control error (ACE) to be minimized. Then the power settings are computed for the generation allocation logic of the generators participating in AGC. In practice, the data to form the ACE signal are usually transmitted through unprotected channels [33], [34]. Thus in this article, we mainly consider an anomaly scenario where FDI attacks are corrupting the frequencies and power exchanges as parts of the ACE signal (and hence the AGC controller dynamics). In particular, we take the instance of stationary FDI attack, i.e., the attack occurs as a constant bias injection $\boldsymbol{f}$ during system operations at a specific time instance and it remains unchanged since then.

An advanced attack attends to pursue a desired impact on the system dynamics and also achieve undetectability from some possible data quality checking programs.[2] Thus, an adversary would try to inject "smart" false data. The next definition opts to formalize this class of attack.

Definition 1 (Disruptive Univariate or Multivariate FDI Attack): Consider a stationary FDI attack with $\boldsymbol{f} \in \mathbb{R}^{n_f}$. In the scenario of univariate attack ($n_f = 1$, only one signal channel is corrupted), we call an FDI attack $\boldsymbol{f} \in \mathcal{F}$ disruptive attack if $\mathcal{F}$ is a set of $\{\boldsymbol{f} \in \mathbb{R}^{n_f=1}\colon \boldsymbol{f}_{min} \le \boldsymbol{f} \le \boldsymbol{f}_{max}\}$ where $\boldsymbol{f}_{min}, \boldsymbol{f}_{max} \in \mathbb{R}^{n_f=1}$ are non-zero variables. We call the set $\mathcal{F}$ plausible as it reflects the disruptive attack's targets on attack impact and undetectability. Similarly, in the scenario of multivariate attack ($n_f > 1$, multiple signal channels are corrupted), we introduce the plausible set as

$$\mathcal{F} = \left\{\boldsymbol{f} \in \mathbb{R}^{n_f} : \boldsymbol{f} = \boldsymbol{F}_{\mathrm{b}}^{\top}\boldsymbol{\alpha},\ \boldsymbol{\alpha} \in \mathcal{A}\right\}$$

where $\boldsymbol{F}_{\mathrm{b}} := [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_d]$ represents a finite basis for the set of multivariate attacks, and $\boldsymbol{f}_i \in \mathbb{R}^{n_f}$ for $i \in \{1, \ldots, d\}$, and $\boldsymbol{\alpha} := [\alpha_1, \alpha_2, \ldots, \alpha_d]^{\top} \in \mathbb{R}^d$ contains the coefficients. $\mathcal{A} := \{\boldsymbol{\alpha} \in \mathbb{R}^d \mid \boldsymbol{A}\boldsymbol{\alpha} \ge \boldsymbol{b}\}$ with $\boldsymbol{A} \in \mathbb{R}^{n_b\times d}$ and $\boldsymbol{b} \in \mathbb{R}^{n_b}$ is polytopic to reflect attack targets on attack impact and undetectability. We emphasize that $\mathcal{F}$ can be adjusted according to different anomaly scenarios where the convexity of the set is particularly desired from a computational perspective in the subsequent analysis.

The disruptive univariate and multivariate attacks to be detected are modeled and implemented in the PowerFactory simulator. In the Appendix, we present the modeling of

[2] For instance, to avoid triggering data quality alarms, generally the calculated ACE in an area should not exceed a permitted value [35], [36].

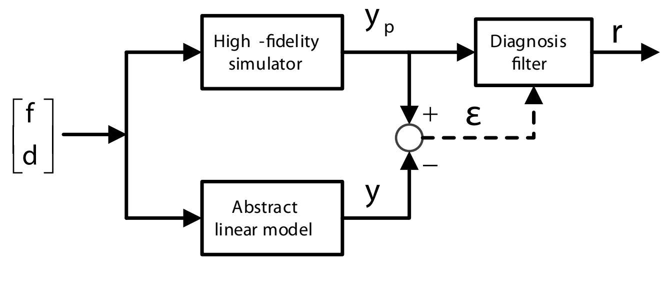


Fig. 2. Configuration of the proposed solution.

univariate and multivariate FDI attacks against AGC functions in PowerFactory. Further details of system modeling in PowerFactory and an instance of the abstract model (2), for the part of Section II-A, are also given.

C. Mathematical Framework of the Proposed Solution

To alleviate the impact of model mismatch, we propose our solution outlined in Figure 2. We denote the system output from the simulation results of the high-fidelity simulator (PowerFactory in this case) as $\boldsymbol{y}_p$. Our proposed solution builds on a new perspective that the diagnosis filter utilizes not only the abstract model-based information for a scalable design but also the simulation data to "train" the filter to achieve performance robustness with respect to model mismatch. Note that model mismatch is reflected through the difference of $\boldsymbol{y}_p$ from the high-fidelity simulator and $\boldsymbol{y}$ from the abstract model (2); such mismatch signature is characterized by $\boldsymbol{\varepsilon}$ in Figure 2.

Let us start with the first approach that builds on the model-based information. In a realistic framework, the output data are applied to the diagnosis filter in discrete-time samples. Thus firstly, we would like to express the discrete-time version of (2) as a more compact linear DAE (here it refers to difference algebraic equation) formulation,

$$\boldsymbol{H}(q)\bar{\boldsymbol{x}}[k] + \boldsymbol{L}(q)\boldsymbol{y}[k] + \boldsymbol{F}(q)\boldsymbol{f}[k] = 0, \ \forall k \in \mathbb{N}, \tag{3}$$

where q is a time-shift operator such that $q\tilde{\boldsymbol{x}}[k] \to \tilde{\boldsymbol{x}}[k+1]$. The augmented vector $\bar{\boldsymbol{x}} := [\tilde{\boldsymbol{x}}^\top \ \boldsymbol{d}^\top]^\top$ consists of both the state variables and the load disturbances. We denote n_r as the number of rows in (3). Then $\boldsymbol{H}$, $\boldsymbol{L}$, $\boldsymbol{F}$ are polynomial matrices in terms of the operator q with n_r rows and n_x, n_y, n_f columns, respectively, by defining,

$$\boldsymbol{H}(q) := \begin{bmatrix} -q\boldsymbol{I} + \boldsymbol{A}_x & \boldsymbol{B}_d \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix}, \boldsymbol{L}(q) := \begin{bmatrix} \boldsymbol{0} \\ -\boldsymbol{I} \end{bmatrix}, \boldsymbol{F}(q) := \begin{bmatrix} \boldsymbol{B}_f \\ \boldsymbol{D}_f \end{bmatrix},$$

where $\boldsymbol{A}_x$, $\boldsymbol{B}_d$ and $\boldsymbol{B}_f$ are from a zero-order hold (ZOH) discretization of (2) for a given sampling time T_s [37, p 314]. Here we clarify that T_s is chosen based on the timescales of interest. In our motivating case study, we consider the AGC process that performs as the secondary frequency control of a multi-area system and has relatively slow dynamics (the timescale of secondary frequency response is in the range of seconds) [31]. We note that T_s differs from the simulation time step set in the high-fidelity simulator whose time step resolution can be very high for power system dynamics simulation. Thus in the following, the simulation results from PowerFactory initially with a very small time step also need to be "sampled" first according to T_s.

Next, we further explain the setting of our proposed solution. For the filter design in Figure 2, $\boldsymbol{y}_p$ from the simulation results of PowerFactory is available as the input to the diagnosis filter. In this article, we propose a diagnosis filter as a type of residual generator having a linear transfer operation,

$$r[k] := \boldsymbol{R}_\varepsilon(q)\boldsymbol{y}_p[k], \tag{4}$$

$$\boldsymbol{\varepsilon}[k] = \boldsymbol{y}_p[k] - \boldsymbol{y}[k], \tag{5}$$

where r is the residual signal (generally a one-dimension signal for the sake of diagnosis) in Figure 2, $\boldsymbol{R}_\varepsilon(q)$ with a predefined degree is the design variable of the diagnosis filter, which will depend on the information of the abstract model in the preceding subsection and also the simulation data through the mismatch signature $\boldsymbol{\varepsilon}$ introduced in (5).

III. Optimization-Based Characterization of the Diagnosis Filter

A. Robust Diagnosis Filter

Considering the model-based information in the compact formulation of (3) from (2), the residual generator can be represented through polynomial matrix equations. Thus for $\boldsymbol{R}_\varepsilon(q)$ in (4), we introduce $\boldsymbol{R}_\varepsilon(q) := a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{L}(q)$, where $\boldsymbol{N}(q) := \sum_{i=1}^{d_N} \boldsymbol{N}_i q^i$ in which $\boldsymbol{N}_i \in \mathbb{R}^{1\times n_r}$ (note that the dimension of the residual signal is 1) for $i \in \{1, \ldots, d_N\}$ and d_N is the predefined degree. Now $\boldsymbol{N}(q)$ becomes the filter design variable, if the scalar polynomial $a(q)$ with sufficient order to make $\boldsymbol{R}_\varepsilon(q)$ physically realizable is determined. Note that d_N is adjustable to be much less than the order of system dynamics in (2). From (3) to (5), we can further have

$$\begin{aligned} r[k] &= a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{L}(q)\boldsymbol{y}_p[k] \\ &= a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{L}(q)(\boldsymbol{y}[k] + \boldsymbol{\varepsilon}[k]) \\ &= -\underbrace{a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{H}(q)\bar{\boldsymbol{x}}[k]}_{\text{(I)}} - \underbrace{a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{F}(q)\boldsymbol{f}[k]}_{\text{(II)}} \\ &\quad + \underbrace{a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{L}(q)\boldsymbol{\varepsilon}[k]}_{\text{(III)}} \end{aligned} \tag{6}$$

where term (II) is the only desired contribution from the anomaly $\boldsymbol{f}$. Ideally, we would like to let the residual keep robust against the unknowns of term (I) and (III). For that purpose, first we need to quantify the effect of model mismatch (reflected through the mismatch signature $\boldsymbol{\varepsilon}$) on the filter residual. For all $k \in \mathbb{N}$, let us define

$$r_\varepsilon[k] := a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{L}(q)\boldsymbol{\varepsilon}[k]. \tag{7}$$

Next, let us further denote the space of a discrete-time signal taking values in $\mathbb{R}^n$ over the horizon of T (i.e., $k \in \{1, \ldots, T\}$) by $\mathcal{W}_T^n$. We equip this space with an inner product and a corresponding norm as

$$\|\boldsymbol{v}\|_{\mathcal{L}_2}^2 := \langle \boldsymbol{v}, \ \boldsymbol{v} \rangle, \quad \langle \boldsymbol{v}, \ \boldsymbol{w} \rangle := \sum_{k=1}^{T} \boldsymbol{v}^\top[k]\boldsymbol{w}[k], \tag{8}$$

where $\boldsymbol{v}$, $\boldsymbol{w}$ are some elements in the space $\mathcal{W}_T^n$.

The main objective of applying a diagnosis filter in the high-fidelity simulator for anomaly detection is to make the filter residual r as sensitive to anomaly $\boldsymbol{f}$ as possible and simultaneously as robust as possible against other unknowns in term (I) and (III). To achieve that, we introduce a scalable and robust diagnosis filter characterized by a class of residual generator that has the following features.

Definition 2 (Robust Diagnosis Filter): Consider the residual generator represented via a polynomial vector $\boldsymbol{N}(q)$ for a given $a(q)$ in (6). This residual generator is robust with respect to model mismatch and can detect all the plausible disruptive attacks, if $\boldsymbol{N}(q)$ is the optimal solution from

$$\begin{aligned} \min_{\boldsymbol{N}(q)} \quad & \|r_\varepsilon\|^2_{\mathcal{L}_2} \\ \text{s.t.} \quad & \boldsymbol{N}(q)\boldsymbol{H}(q) = \boldsymbol{0} \\ & \boldsymbol{N}(q)\boldsymbol{F}(q)\boldsymbol{f} \neq \boldsymbol{0}, \quad \forall \boldsymbol{f} \in \mathcal{F}. \end{aligned} \tag{9}$$

The first constraint in (9) ensures rejection of the unknown of term (I) in the filter residual; the second constraint guarantees the filter sensitivity to all the admissible disruptive FDI attacks in the plausible set $\mathcal{F}$ of Definition 1; the objective function seeks to reduce the impact of model mismatch on the filter residual (term (III)).

B. Tractable Optimization-Based Characterization Under Univariate Attack

In light of (9) in Definition 2 for the robust diagnosis filter design, let us first consider the univariate attack scenario ($n_f = 1$). Note that when there is no attack, the system output $\boldsymbol{y}_p$ from the PowerFactory simulations and $\boldsymbol{y}$ from the abstract model (2) only depend on the input of load disturbances $\boldsymbol{d}$. Thus for one instance of load disturbances, $\boldsymbol{d}_i$, one can have a specific mismatch signature $\boldsymbol{\varepsilon}_i$ according to (5). For each $\boldsymbol{\varepsilon}_i \in \mathcal{W}_T^{n_y}$, a matrix $\boldsymbol{E}_i \in \mathbb{R}^{n_y \times T}$ can be introduced,

$$\boldsymbol{E}_i := [\boldsymbol{\varepsilon}_i[1],\ \boldsymbol{\varepsilon}_i[2],\ \ldots,\ \boldsymbol{\varepsilon}_i[T]]. \tag{10}$$

Recall that the operator q acts as a time-shift operator: $q\boldsymbol{\varepsilon}_i[k] \rightarrow \boldsymbol{\varepsilon}_i[k+1]$. This operator is linear, and it can be translated as a matrix left-shift operator for matrix $\boldsymbol{E}_i$: $q\boldsymbol{E}_i = \boldsymbol{E}_i\boldsymbol{D}$ where $\boldsymbol{D}$ is a square matrix of order T. Following the definition of the residual r_ε in (7), we have

$$a(q)r_\varepsilon = \boldsymbol{N}(q)\boldsymbol{L}(q)\boldsymbol{E}_i = \bar{\boldsymbol{N}}\bar{\boldsymbol{L}}\begin{bmatrix} \boldsymbol{I} \\ q\boldsymbol{I} \\ \vdots \\ q^{d_N}\boldsymbol{I} \end{bmatrix}\boldsymbol{E}_i = \bar{\boldsymbol{N}}\bar{\boldsymbol{L}}\boldsymbol{D}_i \tag{11}$$

where the matrices are defined as $\bar{\boldsymbol{N}} := [\boldsymbol{N}_0, \boldsymbol{N}_1, \ldots, \boldsymbol{N}_{d_N}]$, $\bar{\boldsymbol{L}} := diag[\boldsymbol{L}, \boldsymbol{L}, \ldots, \boldsymbol{L}]$ and $\boldsymbol{L} = \boldsymbol{L}(q)$, and $\boldsymbol{D}_i := [\boldsymbol{E}_i^T,\ (\boldsymbol{E}_i\boldsymbol{D})^T,\ \ldots,\ (\boldsymbol{E}_i\boldsymbol{D}^{d_N})^T]^T$. Given a particular disturbance pattern $\boldsymbol{d}_i$, then the $\mathcal{L}_2$-norm of the residual signal as defined in (8) can be reformulated as a quadratic function,

$$\|r_{\varepsilon_i}\|^2_{\mathcal{L}_2} = \bar{\boldsymbol{N}}\boldsymbol{Q}_i\bar{\boldsymbol{N}}^\top, \quad \boldsymbol{Q}_i = (\bar{\boldsymbol{L}}\boldsymbol{D}_i)\boldsymbol{G}\big(\bar{\boldsymbol{L}}\boldsymbol{D}_i\big)^\top, \tag{12}$$

where $\boldsymbol{G}$ is a positive semi-definite matrix with a dimension of T such that $\boldsymbol{G}(i,j) = \langle a(q)^{-1}u_i,\ a(q)^{-1}u_j\rangle$ in which $u_i, u_j \in \mathcal{W}_T^1$ are the discrete-time unit impulses. It can be observed from (12) that the matrix $\boldsymbol{Q}_i$ is also positive semi-definite since $\boldsymbol{Q}_i$ is symmetric and for all non-zero row vector $\bar{\boldsymbol{N}}$, we can have $\bar{\boldsymbol{N}}\boldsymbol{Q}_i\bar{\boldsymbol{N}}^\top = \|r_{\varepsilon_i}\|^2_{\mathcal{L}_2} \geq 0$. We call $\boldsymbol{Q}_i$ the *mismatch signature matrix* resulting from a specific mismatch signature under a particular load disturbance instance.

Remark 2 (Training With Multiple Model Mismatch Signatures): In order to robustify the diagnosis filter, it can be *"trained"* by utilizing the information of multiple instances of load disturbances, i.e., $\{\boldsymbol{d}_i\}_{i=1}^m$, under normal system operations (without attacks). For each disturbance signature $\boldsymbol{d}_i$, the mismatch signature $\boldsymbol{\varepsilon}_i$ and also the mismatch signature matrices $\boldsymbol{Q}_i$ can be computed from (10) to (12). Next, according to (9) in Definition 2, the robust diagnosis filter has an optimization-based characterization where the objective function can be formulated to minimize $\bar{\boldsymbol{N}}((1/m)\sum_{i=1}^m \boldsymbol{Q}_i)\bar{\boldsymbol{N}}^\top$ (average-cost viewpoint) or $\max_{i\leq m}(\bar{\boldsymbol{N}}\boldsymbol{Q}_i\bar{\boldsymbol{N}}^\top)$ (worst-case viewpoint). We note that from computational perspective the average-cost is much more preferred. It is worthy mentioning here that the training number m may have a significant impact on the diagnosis performance of the proposed filter. As a general rule, the smaller the training number m, the worse the proposed filter may behave in detecting anomalies by performing more misdiagnoses.

Theorem 1 (Tractable Quadratic Programming Characterization): Consider the polynomial matrices $\boldsymbol{H}(q) = \boldsymbol{H}_0 + q\boldsymbol{H}_1$ and $\boldsymbol{F}(q) = \boldsymbol{F}$ where $\boldsymbol{H}_0, \boldsymbol{H}_1 \in \mathbb{R}^{n_r \times n_{\bar{x}}}$ and $\boldsymbol{F} \in \mathbb{R}^{n_r \times n_f}$ are constant matrices. The robust diagnosis filter introduced in (9) of Definition 2 for the univariate attack can be obtained by solving the optimization program,

$$\begin{aligned} \min_{\bar{\boldsymbol{N}}} \quad & \bar{\boldsymbol{N}}\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{Q}_i\right)\bar{\boldsymbol{N}}^\top \\ \text{s.t.} \quad & \bar{\boldsymbol{N}}\bar{\boldsymbol{H}} = \boldsymbol{0} \\ & \big\|\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\big\|_\infty \geq 1 \end{aligned} \tag{13}$$

where $\|\cdot\|_\infty$ denotes the infinite vector norm, and

$$\bar{\boldsymbol{H}} := \begin{bmatrix} \boldsymbol{H}_0 & \boldsymbol{H}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_0 & \boldsymbol{H}_1 & \boldsymbol{0} & \vdots \\ \vdots & \boldsymbol{0} & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{H}_0 & \boldsymbol{H}_1 \end{bmatrix}.$$

Similar to the matrix $\bar{\boldsymbol{L}}$ in (12), $\bar{\boldsymbol{F}}$ in (13) is defined as $\bar{\boldsymbol{F}} := diag[\boldsymbol{F},\ \boldsymbol{F},\ \ldots,\ \boldsymbol{F}]$. Besides, a robust diagnosis filter from the program (13) but simply adding the following linear constraint can have non-zero steady-state residual that could approximate the attack value of $\boldsymbol{f}$,

$$-a(1)^{-1}\sum_{i=0}^{d_N}\boldsymbol{N}_i\boldsymbol{F} = 1. \tag{14}$$

Proof: The key step is to observe that in (9) we can rewrite,

$$\begin{aligned} \boldsymbol{N}(q)\boldsymbol{H}(q) &= \bar{\boldsymbol{N}}\bar{\boldsymbol{H}}\Big[\boldsymbol{I},\ q\boldsymbol{I},\ \ldots,\ q^{d_N+1}\boldsymbol{I}\Big]^\top, \\ \boldsymbol{N}(q)\boldsymbol{F}(q) &= \bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\Big[\boldsymbol{I},\ q\boldsymbol{I},\ \ldots,\ q^{d_N}\boldsymbol{I}\Big]^\top. \end{aligned} \tag{15}$$

In the scenario of univariate attack ($n_f = 1$), the last constraint of (9) can be translated into $\boldsymbol{N}(q)\boldsymbol{F}(q) \neq 0$ to ensure a nonzero transfer function from the univariate attack to the residual signal (i.e., a non-zero response when the univariate attack occurs). Then looking at the linear structure of $\boldsymbol{N}(q)\boldsymbol{F}(q) \neq 0$, from (15), one can scale this inequality to arrive at the one of $\|\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\|_\infty \geq 1$ in (13). Next, the constraint $\boldsymbol{N}(q)\boldsymbol{H}(q) = 0$ in (9) can be recast as $\bar{\boldsymbol{N}}\bar{\boldsymbol{H}} = 0$ in (13), following (15). Then, when $\bar{\boldsymbol{N}}\bar{\boldsymbol{H}} = 0$ satisfies, the diagnosis filter becomes $r[k] = -a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{F}(q)\boldsymbol{f}[k] + a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{L}(q)\boldsymbol{\varepsilon}[k]$. If there is no model mismatch ($\boldsymbol{\varepsilon} \equiv \boldsymbol{0}$), as a discrete-time signal, the steady-state value of the filter residual under the univariate attack would be $-a(q)^{-1}\boldsymbol{N}(q)\boldsymbol{F}(q)\boldsymbol{f}|_{q=1}$. Note that $\boldsymbol{N}(1)\boldsymbol{F}(1) = \sum_{i=0}^{d_N}\boldsymbol{N}_i\boldsymbol{F}$. Thus when there exists model mismatch, with the linear constraint of (14), the residual in its steady-state value could approximate $\boldsymbol{f}$. ■

For the last constraint of (13), as indicated by [17, Lemma 4.3], $\|\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\|_\infty \geq 1$ if and only if there exists a coordinate j that $\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\boldsymbol{v}_j \geq 1$ or $\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\boldsymbol{v}_j \leq -1$. Here $\boldsymbol{v}_j = [0, \ldots, 1, \ldots, 0]^T_{(d_N+1)}$ in which the only non-zero element of the vector is the j-th element. Thus, one can view (13) as a family of $d_N + 1$ standard QPs with linear constraints that for each QP the last constraint becomes $\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\boldsymbol{v}_j \geq 1$ (or $\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\boldsymbol{v}_j \leq -1$ since the set for the solutions of $\bar{\boldsymbol{N}}$ in (13) is symmetric). In addition, recall that the matrix $\boldsymbol{Q}_i$ in the objective function is positive semi-definite, which implies that the resulted standard QPs are also convex, and hence tractable.

Remark 3 (Computational Complexity): For the family of $d_N + 1$ convex QPs, each of them contains 2 constraints and $n_r(d_N + 1)$ decision variables (recall the dimension of $\bar{\boldsymbol{N}}$). Here n_r, as the number of rows in the compact model (3), depends on the number of state variables and system outputs of the picked abstract model (2) (i.e., $n_r = n_{\tilde{x}} + n_y$). In our motivating case study of Section II, such number is indeed determined by the abstract model of the studied power system (e.g., the number of areas, transmission lines called tie-lines that connect areas, and generators participating in AGC). But, notably, another parameter d_N, as the degree of the diagnosis filter, also affects the number of resulted QPs and decision variables. The value of d_N is adjustable and can be much less then the dimension of system dynamics. That is to say, our optimization-based characterization for the proposed diagnosis filter results in highly scalable optimization programs. Besides, a convex QP can be solved in polynomial time and there are very efficient toolboxes (e.g., CPLEX, MOSEK) for that purpose [38]. Thus, considering that the training phase of our approach is a matter of matrix computation and the QPs are convex, we would say that our approach is not computational expensive especially comparing with many machine-learning type methods, but scalable for different studied systems.

For a better illustration, Algorithm 1 concludes the diagnosis filter construction and validation process for an implementation in PowerFactory to detect the univariate attack.

C. Extension to Multivariate Attack Scenarios

Inspired by the techniques developed in [5], we further extend the preceding design of the robust diagnosis filter to the scenario of multivariate attacks.

Algorithm 1 Diagnosis Filter Validation for Univariate Attack

```
(i) Training phase
    Input: $y_p$, $y$ under normal operations, $d_N$, $a(q)$
    Output: $\bar{N}$, $R_\varepsilon(q)$
      1 For each $d_i$, from $y_p$ and $y$, compute the mismatch
        signature matrix $Q_i$ according to (10) - (12).
      2 For a number of $m$ instances of $d_i$, perform the first
        step to get $m$ signature matrices.
      3 Solve the family of convex QPs from (13) with the
        derived $Q_i$. Build $R_\varepsilon(q)$ based on the solution $\bar{N}$
        from solving QPs and also $d_N$, $a(q)$.
(ii) Testing phase:
    Input: $y_p$ under univariate attacks, $R_\varepsilon(q)$
    Output: $r$ for anomaly detection
      1 Let $y_p$ be the input of the diagnosis filter with
        $R_\varepsilon(q)$. Check the residual $r$ for anomaly detection.
```

Corollary 1 (Robust Diagnosis Filter Under Multivariate Attacks): Consider the diagnosis filter in Definition 2 where the set of multivariate attacks is defined as $\mathcal{F} = \{\boldsymbol{f} \in \mathbb{R}^{n_f} : \boldsymbol{f} = \boldsymbol{F}_{\mathrm{b}}^\top\boldsymbol{\alpha},\ \boldsymbol{\alpha} \in \mathcal{A}\}$ in which $\mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}^d \mid \boldsymbol{A}\boldsymbol{\alpha} \geq \boldsymbol{b}\}$ (see Definition 2 for the denotation of these variables). Given $j \in \{1, \ldots, 2d_N + 2\}$, for each j, consider a family of the following quadratic programs,

$$\begin{aligned} \min_{\bar{\boldsymbol{N}},\boldsymbol{\lambda}}\ & \bar{\boldsymbol{N}}\left(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{Q}_i\right)\bar{\boldsymbol{N}}^\top \\ \text{s.t.}\ & \boldsymbol{b}^\top\boldsymbol{\lambda} \geq \gamma_j, \\ & (-1)^j\boldsymbol{N}_{\lfloor j/2 \rfloor}\boldsymbol{F}\boldsymbol{F}_{\mathrm{b}} = \boldsymbol{\lambda}^\top\boldsymbol{A} \\ & \bar{\boldsymbol{N}}\bar{\boldsymbol{H}} = \boldsymbol{0},\ \boldsymbol{\lambda} \geq \boldsymbol{0} \end{aligned} \tag{QP$_j$}$$

where $\lfloor\cdot\rfloor$ is the ceiling function that maps the argument to the least integer. Then, the best solution of the quadratic programs (QP$_j$) among $j \in \{1, \ldots, 2d_N+2\}$ solve the problem (9) in Definition 2 of robust diagnosis filter for the scenario of multivariate attacks.

Proof: In the scenario of multivariate attacks, the two constraints in (9) can be characterized by the maximin program,

$$\gamma^\star := \max_{\bar{\boldsymbol{N}}\in\mathcal{N}}\ \min_{\boldsymbol{\alpha}\in\mathcal{A}}\ \big\{\mathcal{J}(\bar{\boldsymbol{N}}, \boldsymbol{\alpha})\big\}, \tag{16}$$

where the set $\mathcal{N} := \{\bar{\boldsymbol{N}} \in \mathbb{R}^{(d_N+1)n_r} \mid \bar{\boldsymbol{N}}\bar{\boldsymbol{H}} = \boldsymbol{0}\}$. The source of the cost function $\mathcal{J}(\bar{\boldsymbol{N}}, \boldsymbol{\alpha})$ is referred to [5, Sec. IV.B]. Then, according to [5, Th. IV.3], we know that the maximin program (16) can be reformulated and relaxed to a set of linear programs (LPs),

$$\begin{aligned} \gamma_j^\star := \max_{\bar{\boldsymbol{N}},\boldsymbol{\lambda}}\ & \boldsymbol{b}^\top\boldsymbol{\lambda} \\ \text{s.t.}\ & (-1)^j\boldsymbol{N}_{\lfloor j/2 \rfloor}\boldsymbol{F}\boldsymbol{F}_{\mathrm{b}} = \boldsymbol{\lambda}^\top\boldsymbol{A}, \\ & \bar{\boldsymbol{N}}\bar{\boldsymbol{H}} = \boldsymbol{0},\ \boldsymbol{\lambda} \geq \boldsymbol{0} \end{aligned} \tag{LP$_j$}$$

Namely, the solution to the program (LP$_j$) is a feasible solution to the maximin program (16), and $\max_{\{j\leq 2d_N+2\}}\gamma_j^\star \leq \gamma^\star$. Then it is easy to obtain the finite (QP$_j$) for the multivariate attack scenario. We conclude the proof by noting that if there is a $\gamma_j^\star$

Algorithm 2 Diagnosis Filter Validation for Multivariate Attack

```
(i) Pre-training
    Input: $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{F}_{\mathrm{b}}$
    Output: $\gamma_j^\star$
        1 Solve (LP$_j$) for each $j \in \{1, \ldots, 2d_N + 2\}$. Check
          if there exists $\gamma_j^\star > 0$ and find the maximum.
(i) Training phase
    Input: $\boldsymbol{y}_p$, $\boldsymbol{y}$ under normal operations, $d_N$, $a(q)$
    Output: $\bar{\boldsymbol{N}}$, $\boldsymbol{R}_\varepsilon(q)$
        1 For each $\boldsymbol{d}_i$, from $\boldsymbol{y}_p$ and $\boldsymbol{y}$, compute the mismatch
          signature matrix $\boldsymbol{Q}_i$ according to (10) - (12).
        2 For a number of $m$ instances of $\boldsymbol{d}_i$, perform the first
          step to get $m$ signature matrices.
        3 Set the initial value of $\gamma_j$ to be $\max_{\{j \le 2d_N+2\}} \gamma_j^\star$
          from pre-training. Solve (QP$_j$) with the derived $\boldsymbol{Q}_i$.
        4 Tune the value of $\gamma_j$ until it reaches maximum.
          Build $\boldsymbol{R}_\varepsilon(q)$ based on the solution $\bar{\boldsymbol{N}}$ and $d_N$, $a(q)$.
(ii) Testing phase:
    Input: $\boldsymbol{y}_p$ under multivariate attacks, $\boldsymbol{R}_\varepsilon(q)$
    Output: $r$ for anomaly detection
        1 Let $\boldsymbol{y}_p$ be the input of the diagnosis filter with
          $\boldsymbol{R}_\varepsilon(q)$. Check the residual $r$ for anomaly detection.
```

being positive,[3] then a resulted filter from (LP$_j$) could detect all the multivariate attacks in the set $\mathcal{F}$. ■

From Corollary 1, we can see that for any $j \in \{1, \ldots, 2d_N + 2\}$, if one can find a $\gamma_j > 0$ that (QP$_j$) is still feasible, then the solution to QP$_j$ offers a robust diagnosis filter in the type of Definition 2 for multivariate attacks. It is worth mentioning that the obtained (QP$_j$) in Corollary 1 is also a standard convex QP. Along the same lines as Remark 3, each program of (QP$_j$) ($j \in \{1, \ldots, 2d_N + 2\}$) in Corollary 1 has $n_r(d_N + 1) + n_b$ decision variables and can be solved in polynomial time. Thus it can be observed that for the scenario of multivariate attacks, our proposed approach can also achieve high scalability.

Algorithm 2 concludes the filter construction and validation process of our proposed solution in the scenario of multivariate attacks. In the "pre-training", one needs to solve (LP$_j$) for each j to see if there exists $\gamma_j^\star > 0$. If yes, next in the "training phase", similar to the process in Algorithm 1, the mismatch signature matrices can be computed according to (10) - (12). Then the program (QP$_j$) needs to be solved and the resulted robust diagnosis filter can be "tested" in PowerFactory. We would like to highlight that, the robust diagnosis filter from (QP$_j$) does not necessarily enforce a non-zero steady-state residual under multivariate attacks. Regarding its steady-state behavior, the program (16) can be modified into $\mu^\star := \max_{\{\bar{N} \in \mathcal{N}\}} \min_{\{\boldsymbol{\alpha} \in \mathcal{A}\}} |\bar{\boldsymbol{N}}\bar{\boldsymbol{F}}\boldsymbol{\alpha}|$ which has an exact convex reformulation. Then a similar treatment as the one in Corollary 1 can be deployed.

TABLE I
A SUMMARY OF FILTER FEATURES IN DIFFERENT SCENARIOS

	Attack scenario		**Model-based detection**	
	univariate	multivariate	pure	data-assisted
Filter features	$n_f = 1$	$n_f > 1$	$\boldsymbol{Q}_i = \boldsymbol{0}$, in [5]	$\boldsymbol{Q}_i$ from (12), this study

In the end, we summarize the filter features under univariate or multivariate attacks and in the pure model-based or our proposed data-assisted model-based methods, in Table I.

IV. Numerical Results

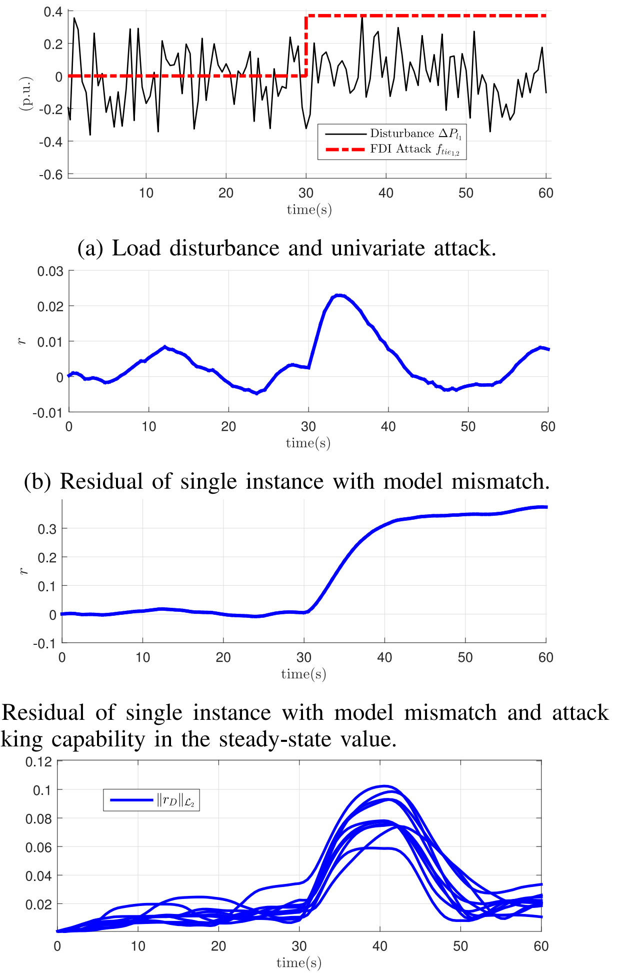
A. Test System and Robust Filter Description

To validate the effectiveness of our data-assisted model-based diagnosis filter, it has been implemented in the high-fidelity simulator of DIgSILENT PowerFactory to detect FDI attacks on the ACE signal of AGC in the three-area 39-bus system. The system parameters of the abstract model are referred to [39], and the specifications of the model in PowerFactory are available at [40]. Following Algorithms 1 and 2, to obtain model mismatch signatures, we run the simulations to obtain $\boldsymbol{y}_p$ and $\boldsymbol{y}$ with the same input $\boldsymbol{d}$ of load disturbances in normal operations. The adjustable degree of the residual generator is set to $d_N = 3$ which is much less than the order of the abstract model (it is a 19-order model of (2)); we set the scalar polynomial $a(q) = (q{-}p)^{d_N}/(1 - p)^{d_N}$ where p is a user-defined variable acting as the pole of $\boldsymbol{R}_\varepsilon(q)$, and it is normalized in steady-state value for all feasible poles. The simulation time step is set to 0.01 s in the RMS simulation tool of PowerFactory. We let $T_s = 0.5$ s such that the simulation results from PowerFactory initially with the very small time step are also "sampled", and thus for a simulation time $t_s = 10$ s, $T = 20$ in (10). We also perform a comparison study for the two methods: our data-assisted model-based filter and the pure model-based one in [5].

B. Main Simulation Results

The first simulation considers the univariate attack where an attacker has manipulated the power exchange between Area 1 and Area 2 from $t = 30$ s in the horizon of 60 s. To challenge the filter, the disturbances are modeled as stochastic load patterns: load variation of Load 4 in Area 1 is a random zero-mean Gaussian signal. A number of $m = 100$ load disturbance instances are generated for the "training phase" where for each disturbance instance $\boldsymbol{d}_i$, a simulation with $t_s = 10$ s is conducted to obtain mismatch signatures $\boldsymbol{\varepsilon}_i$. Following Algorithm 1, the design variable $\bar{\boldsymbol{N}}$ of our diagnosis filter is derived. To compare, a pure model-based filter is also obtained by letting $\boldsymbol{Q}_i = 0$ in Theorem 1 (see Table I), which can be transformed into finite LPs. The simulation results are referred to Figure 3 and Figure 4. We can see that our data-assisted model-based filter has significant improvements in the regards of mitigating the effect from model mismatch on the residual, comparing with the pure model-based approach. Besides, from Figure 4(c), it can track attack value through its

[3] To remark, as noted in Definition 1, the parameter $\boldsymbol{b}$, $\boldsymbol{A}$ and $\boldsymbol{F}_{\mathrm{b}}$ are scenario-specific. There is no need that the elements of $\boldsymbol{b}$ are fully positive; however, if all of them are negative, there may not exist a positive $\gamma_j^\star$.



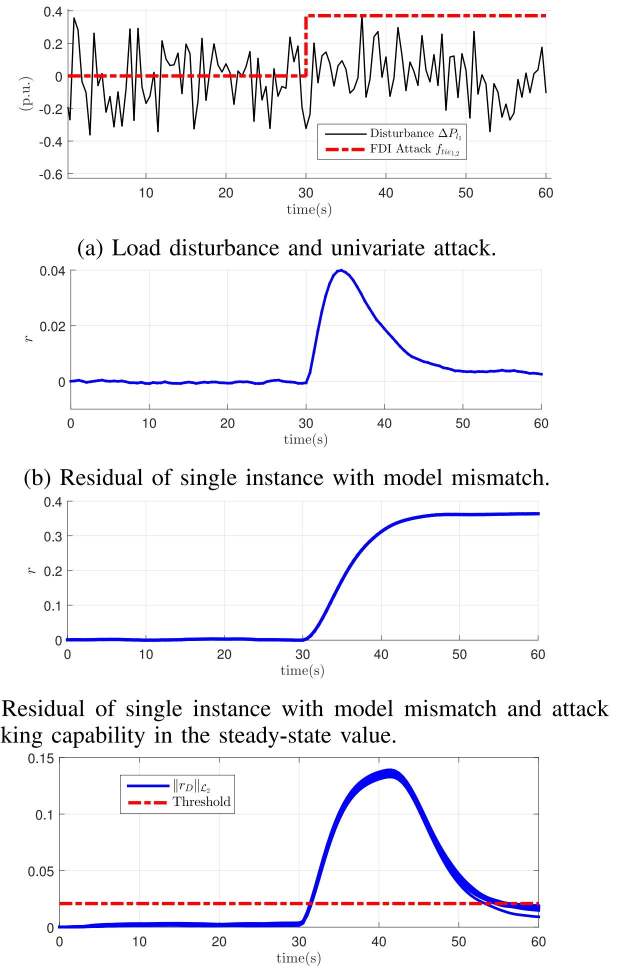
(a) Load disturbance and univariate attack.

(b) Residual of single instance with model mismatch.

(c) Residual of single instance with model mismatch and attack tracking capability in the steady-state value.

(d) Energy of residual of multiple instances with model mismatch.

Fig. 3. Pure model-based filter in [5] under univariate attack. It is derived by letting $\boldsymbol{Q}_i = 0$ in Theorem 1, which can be transformed into finite LPs.



(a) Load disturbance and univariate attack.

(b) Residual of single instance with model mismatch.

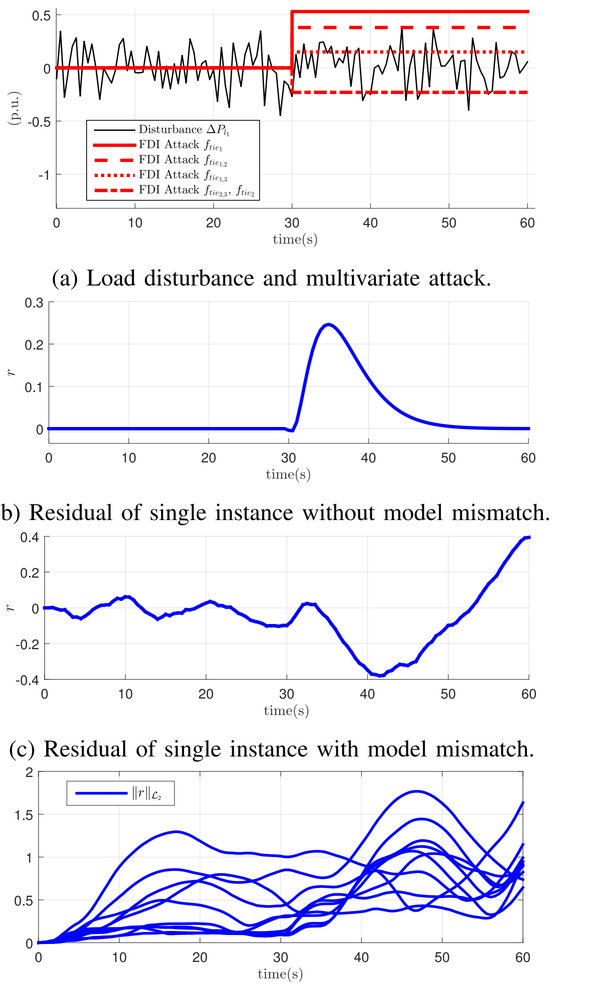
(c) Residual of single instance with model mismatch and attack tracking capability in the steady-state value.

(d) Energy of residual of multiple instances with model mismatch.

Fig. 4. Data-assisted model-based filter. It is derived by solving (13) in Theorem 1 where $\boldsymbol{Q}_i$ is from (12), and in the end it is a family of convex QPs.

steady-state residual. The pure model-based filter fails by triggering false alarms, when applied directly to the PowerFactory simulator. This can be expected since it utilizes the abstract model-based information only and suits for the given model of (2). Figures 3(d) and 4(d) provide the residual results under 10 different realizations of load disturbances in the "testing phase". They depict the "energy" of the residual signals for the last 10s under 10 load disturbance instances, namely $\|r\|_{\mathcal{L}_2}[\cdot]$. Note that in Figure 4(d) the threshold is set to $\tau^\star + 0.025$, where the square of $\tau^\star$ equals to the maximum value of $\bar{\boldsymbol{N}}\boldsymbol{Q}_i\bar{\boldsymbol{N}}$ in the 100 training instances ($i \in \{1, \ldots, 100\}$); recall (12) that $\|r_{\varepsilon_i}\|^2_{\mathcal{L}_2} = \bar{\boldsymbol{N}}\boldsymbol{Q}_i\bar{\boldsymbol{N}}^\top$. The added value is to avoid possible false alarms according to [41]. Then a univariate FDI attack is said to be detected when the value of $\|r\|_{\mathcal{L}_2}[\,\cdot\,]$ is beyond the threshold; we see successful detections by our proposed method in Figure 4(d), while the pure model-based one fails in Figure 3(d). To conclude, our proposed diagnosis filter implemented in the high-fidelity simulator (PowerFactory in our case study) can successfully generate residual "alerts" for the occurrence of FDI attacks, and keep the impact of model mismatch minimized.

In the second simulation, we move to the scenario of multivariate attacks. There are 5 power exchanges between areas that are attacked, and correspondingly there exist 3 basis vectors in the spanning set $\mathcal{F}$: $\boldsymbol{f}_1 = [0.1 \quad 0 \quad 0.1 \quad 0 \quad 0]^T$, $\boldsymbol{f}_2 = [0.1 \quad 0.15 \quad 0.25 \quad 0 \quad 0]^T$, $\boldsymbol{f}_3 = [0 \quad 0 \quad 0 \quad 0.1 \quad 0.1]^T$ (all in p.u.). Besides, for the set of disruptive multivariate attacks, the parameters are set to $\boldsymbol{A} = \boldsymbol{1}^\top$ and $\boldsymbol{b} = [1.5]^\top$ in $\mathcal{A}$. We refer to [5, Sec. V] for the specification of these values. Following Algorithm 2, the program (LP$_j$) is solved first. The optimal value achieves maximum for $j = 2$ that $\gamma_2^\star = 300$, which implies that a diagnosis filter of our approach could be obtained. Next, in the "training phase", a number of $m = 100$ load disturbance instances are randomly generated. The program (QP$_j$) in Corollary 1 is solved for the filter design. For the derived optimal solution $\bar{\boldsymbol{N}}$, the multivariate attack coordinate vector $\boldsymbol{\alpha}$ is obtained by solving the inner minimization of the program (16). In the "test phase", simulations in PowerFactory are conducted that several realizations of load disturbances have been implemented and the multivariate attacks with $\boldsymbol{\alpha}$ have been launched. The performance of the two filters (the filter of our approach and the pure model-based filter from [5]) is validated with two sets of outputs: one from the abstract model (2) (i.e., without model mismatch, $\boldsymbol{\varepsilon} \equiv 0$) and another one from the PowerFactory simulations



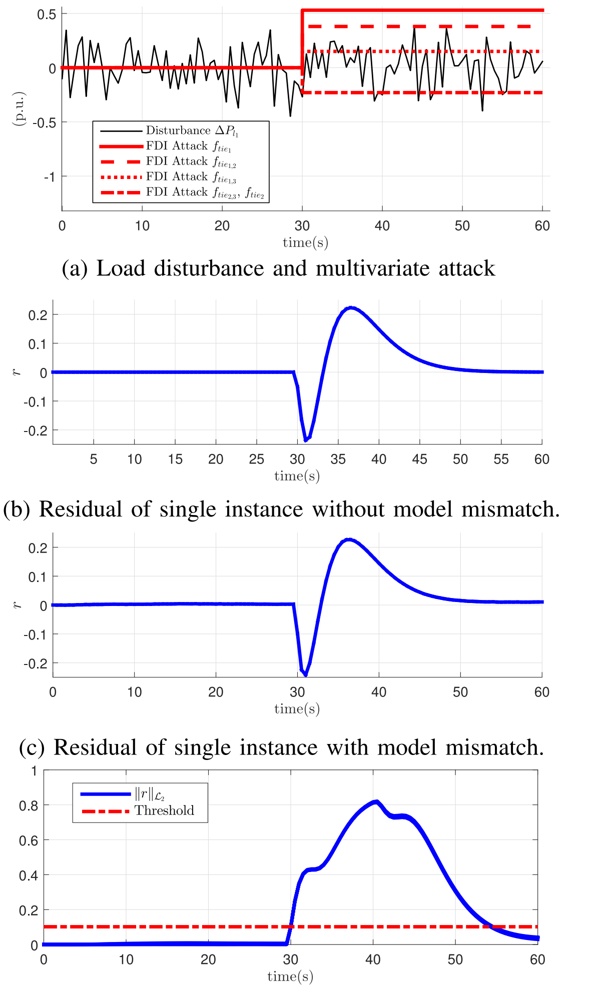
(a) Load disturbance and multivariate attack.

(b) Residual of single instance without model mismatch.

(c) Residual of single instance with model mismatch.

(d) Energy of residual of multiple instances with model mismatch.

Fig. 5. Pure model-based filter in [5]. It is derived by letting $\boldsymbol{Q}_i = 0$ in Corollary 1, and it is essentially a set of LPs.



(a) Load disturbance and multivariate attack

(b) Residual of single instance without model mismatch.

(c) Residual of single instance with model mismatch.

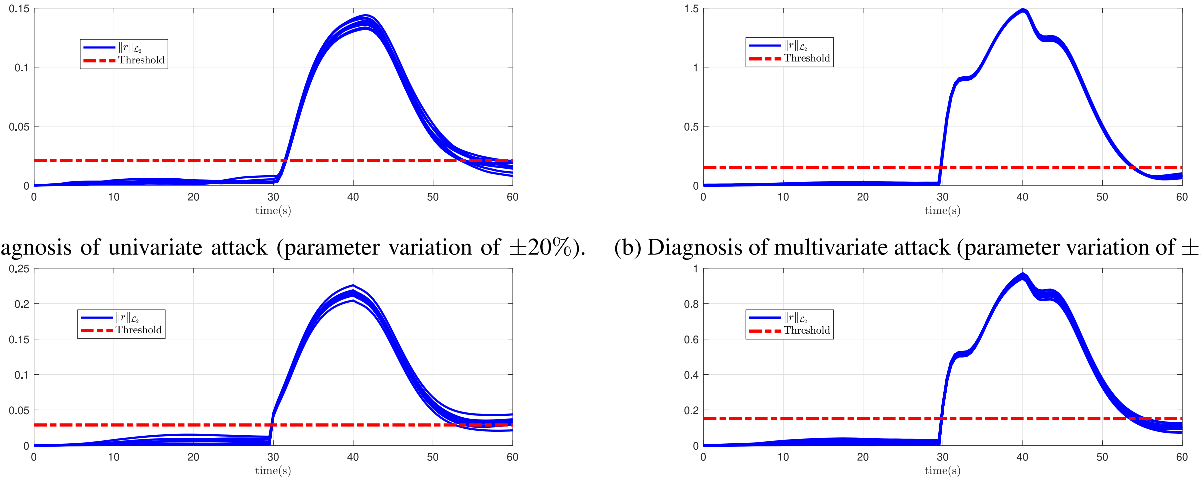
(d) Energy of residual of multiple instances with model mismatch.

Fig. 6. Data-assisted model-based filter. It is derived by solving (QP$_j$) in Corollary 1 where $\boldsymbol{Q}_i$ is from (12).

(i.e., with model mismatch, $\boldsymbol{\varepsilon} \neq 0$). Figures 5 and 6 show the simulation results of both diagnosis filters. We can see that both filters succeed for the case $\boldsymbol{\varepsilon} \equiv 0$. However, from Figures 5(c) and 6(c), when there exists model mismatch, our data-assisted model-based filter still works effectively, while the pure model-based filter in [5] totally fails by triggering false alarms. Besides, similar to Figures 3(d) and 4(d), Figures 5(d) and 6(d) depict the "energy" of the residual signals for the last 10s under 10 load disturbance instances. In Figure 6(d), the threshold is set to $\tau^\star + 0.1$ where the square of $\tau^\star$ equals to the maximum value of $\bar{\boldsymbol{N}}\boldsymbol{Q}_i\bar{\boldsymbol{N}}$ in the 100 training instances ($i \in \{1, \ldots, 100\}$), and the added value is to avoid possible false alarms. Then a multivariate attack is said to be detected when the value of $\|r\|_{\mathcal{L}_2}[\,\cdot\,]$ is beyond this threshold; we see successful detections by our proposed method in Figure 6(d), while the pure model-based one totally fails in Figure 5(d). From Figure 6(d) and Figure 4(d), one can observe that the energy of residual varies slightly for the 10 different instances by our proposed diagnosis filter, compared with the one by the pure model-based method in [5]. The major difference between these two approaches is that we have utilized another important piece of information, the simulation data from PowerFactory to reveal the model mismatch signatures, in addition to the abstract model-based knowledge. On one hand, this phenomenon highlights the importance of considering possible model mismatches when applying model-based diagnosis tool in practice; on the other hand, the results prove the effectiveness of our proposed solution for tackling with that. In the end, note that when looking into the steady-state behavior of the filter, it turns out that $\mu^\star = 0$, which indicates that the optimal multivariate attack in this case is a stealthy attack in the long-term horizon, with or without considering the model mismatch impact. However, one can still detect such attacks with a non-zero transient residual, as shown in Figure 6(d). In conclusion, these simulation results have validated the effectiveness of our proposed solution.

C. Additional Simulation Results and Discussions

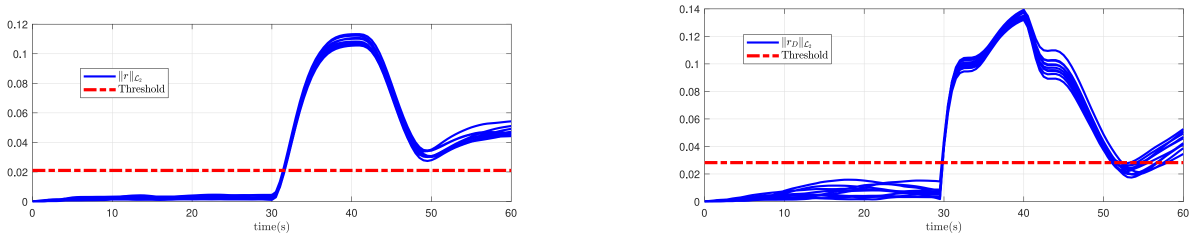
1) Robustness of the Diagnosis Filter to Parameter Variations in the Abstract Model and Measurement Noises: As highlighted in Remark 1, various sources may contribute to the model mismatch between (1) and (2). Thus the following question comes: can the effectiveness of our proposed



(a) Diagnosis of univariate attack (parameter variation of ±20%). (b) Diagnosis of multivariate attack (parameter variation of ±20%).

(c) Diagnosis of univariate attack (parameter variation of ±40%). (d) Diagnosis of multivariate attack (parameter variation of ±40%).

Fig. 7. Data-assisted model-based filter. Variations on parameters (equivalent inertia constants, damping coefficients, droop coefficients) of ±20% and ±40% are made on the abstract model.



(a) Diagnosis of univariate attack under measurement noises. (b) Diagnosis of multivariate attack under measurement noises.

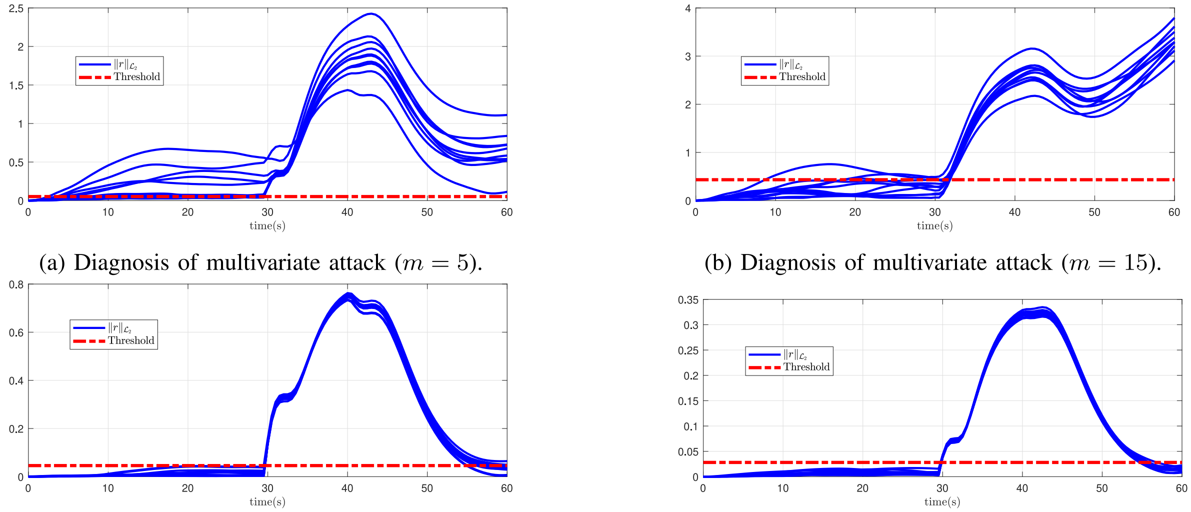
Fig. 8. Data-assisted model-based filter. Measurement noises are considered in the measurements (outputs of simulation results from PowerFactory) in the filter design.

solution on the robustification scheme against model mismatch be still ensured in the cases of other model mismatch sources in addition to load disturbances and cyber attacks? To explore that, we have conducted the following simulations.

First, we would like to explore how the parameter variations in the abstract model could affect the diagnosis performance, which is also a big concern when one needs to chose the abstract model for model-based information in the filter design. Figure 7 shows the energy of residuals under univariate or multivariate attacks, while different levels (±20%, ±40%) of parameter variations on equivalent inertia constants, damping coefficients and droop coefficients of the three areas are considered. We can observe that the proposed diagnosis filter still detects the univariate and multivariate attacks effectively without triggering false alarms when there is no attack (in the period of 0 to 30s in Figure 7), even with parameter variation of ±40% on the abstract model.

The second simulation considers possible measurement noises which are also among the sources contributing to the model mismatch. Here the zero-mean Gaussian noise term added to the measurements (outputs of simulation results from PowerFactory, all in p.u.) follows that the covariance of the frequency measurement is 0.009 and the covariance of other measurements' noise is 0.03 [13]. Then Figure 8 provides the results for both scenarios of univariate and multivariate attacks, when the measurement noises are considered. It can be seen that measurement noises do affect the diagnosis signal; however, we still see a successful detection by our proposed diagnosis filter. But we notice that if the covariance in the noise term becomes much larger (which may be not the case in reality since in general we can have accurate measurements with the development of sensing technology), there may exist the case that our diagnosis filter generates false alarms when there is no attack. This is due to the fact that too much "uncertainty" in the model mismatch may make our robustification scheme fail to capture the "true pattern" of model mismatch signatures. To conclude in this end, Figure 8 and Figure 7 still illustrate the effectiveness of our robustification scheme for possible model mismatches caused by parameter variations and measurement noises.

2) The Impact of Training Number on the Diagnosis Performance: In Remark 2, we have mentioned that the training number may affect the diagnosis performance of our proposed filter. Here we decide to provide more numerical results in order to illustrate how the training number impact the filter residual. To recall first, we have seen that in Section IV-B our diagnosis filter achieves successful detections for both scenarios of univariate and multivariate attacks when the training number is set to 100. In Figure 9, we have conducted simulations under different training numbers in the scenario of multivariate attacks. The results in Figure 9 generally follow the rule noted in Remark 2, i.e., the smaller the training



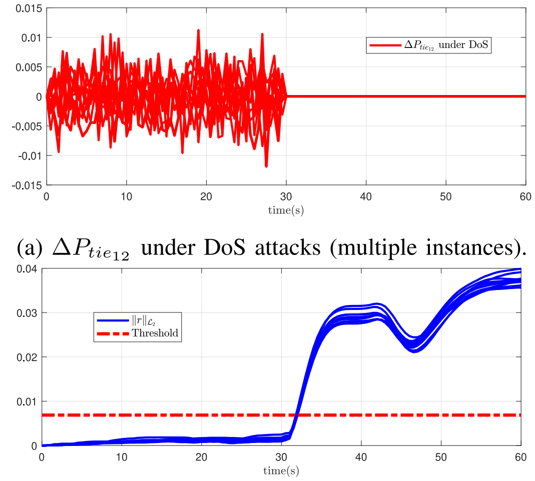
(a) Diagnosis of multivariate attack ($m = 5$).

(b) Diagnosis of multivariate attack ($m = 15$).

(c) Diagnosis of multivariate attack ($m = 35$).

(d) Diagnosis of multivariate attack ($m = 65$).

Fig. 9. Data-assisted model-based filter. The training number varies from $m = 5$ to $m = 65$.



(a) $\Delta P_{tie_{12}}$ under DoS attacks (multiple instances).

(b) Diagnosis of DoS attacks (multiple instances).

Fig. 10. Data-assisted model-based filter in detecting DoS attacks which are launched at 30s, under multiple load disturbance instances.

number *m*, the worse the proposed filter may behave in detecting anomalies by performing more misdiagnoses. In particular, from Figure 9 we see that the diagnosis filter can perform much better when *m* reaches 35 (Figure 9(c)), though there still exist misdiagnoses that false alarms may be triggered. When *m* reaches 65, the diagnosis filter can perform all successful detections in the test set.

3) Detection of Other Types of Cyber Attacks: In this article, the motivating case study comes from the cyber security concerns of the multi-area power system under AGC operations. The FDI attacks on AGC power flow or frequency measurements are essentially acting as exogenous inputs to the system (see $\boldsymbol{f}$ in (2)). There are other types of attacks, e.g., missing data attack (Denial-of-service attack), data repetition attack (replay attack), fault-resembling injection attack, parameter manipulation attack, can be also modeled as the exogenous inputs to the system. Here we introduce two types of them in brief:

- *Denial-of-service (DoS) attack:* A type of missing data attack where the attacker aims to prevent some specific data from being delivered to the respective destinations.
- *Replay attack:* A type of data repetition attack that there exist two stages where the attacker gathers a sequence of data packets at stage 1, and then replays the recorded data afterwards at stage 2.

From a detection point of view, DoS attacks are trivially detectable without any sophisticated mechanisms as the absence of data is not stealthy. In the typical DoS attack modeling, the missing data is typically replaced with the last received ones [42]. In such a mechanism, the DoS can be treated as an "injection" attack. We investigate the performance of our filter in the presence of this type of attacks in Figure 10. Numerical results confirm that our proposed filter can successfully detect the DoS attacks. In regard with the replay attack, the articles [43], [44] offer sufficient conditions under which plausible attacks may remain stealthy irrespective of the detection mechanism providing that the attacker has access all the necessary data channels and excite attack of stage 2 at a suitable time. For instance, if the attacker has obtained all the system outputs under one load disturbance instance and excite stage-2 attack, our proposed diagnosis filter may fail to distinguish the replay attack from load disturbances.

4) Applicability of the Proposed Solution: It can be extended to other systems and anomaly scenarios if the anomalies are still acting as exogenous inputs. Another instance can be the small-signal dynamics model of power systems under anomalies such as cyber attacks or bad data on the measurements (e.g., terminal voltage/current phasor of each generator). The mathematical description of the model can be a linear one considering a small perturbation over an operating point, resulting an abstract model [12, Sec. 3]. We see the effectiveness of our proposed solution with the PowerFactory simulator in our case study, however, we need to clarify that the efficacy of our method is in fact independent on a particular high-fidelity simulator.

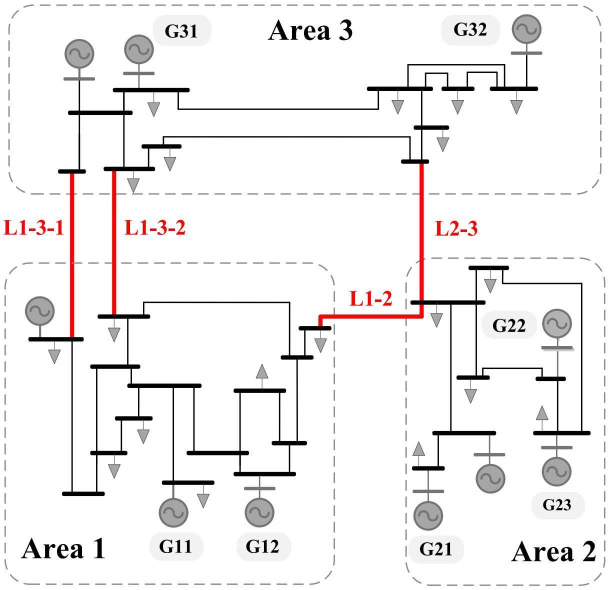


Fig. 11. The diagram of the three-area 39-bus system [5].

5) Comparison With Other Pure Model-Based or Data-Driven Detectors: As illustrated in simulation results, it can be expected that model mismatch would affect the diagnosis of other pure model-based detectors like observers when applied to the high-fidelity simulator. Regarding data-driven approaches, we admit that they may also be able to detect the anomalies of this article, while the data training stage may require a high computational cost, comparing with our proposed tractable convex optimization-based characterization.

V. Conclusion

We have proposed a feasible solution to the problem that arises from applying scalable model-based anomaly detectors in practice: there always exist mismatches between the output from the picked abstract model and the one from the simulator (or the real electric power system). In our final tractable reformulation, the abstract model-based information introduces feasible sets, and the simulation data forms the objective function to minimize the model mismatch effects, which could bridge the model-based and data-driven approaches.

Appendix

A. Parameters of the Abstract Model

An instance of the abstract model (2) is given as follows. Figure 11 depicts the schematic diagram of the three-area 39-bus system used in the abstract model (2) in Section II. Please also note the connection between Figure 11 and Figure 1. For AGC analysis, we are interested in collective performance of all generators, and thus we can rely on certain levels of abstraction that simplify some elements of the initial model in (1) and utilize the possible decoupling between control loops. Then the mathematical description of an AGC system in Area i can be presented in the linear formulation where each area of a power system is represented by a model with equivalent governors, turbines and generators,

$$\begin{cases} \Delta\dot{\omega}_i = \frac{1}{2H_i}\big(\Delta P_{m_i} - \Delta P_{tie_i} - \Delta P_{d_i} - D_i\Delta\omega_i\big), \\ \Delta P_{m_i} = \sum_{g=1}^{G_i} \Delta P_{m_{ig}}, \ \ \Delta P_{tie_i} = \sum_{j\in\mathcal{M}_i} \Delta P_{tie_{ij}}, \\ \Delta\dot{P}_{m_{ig}} = -\frac{1}{T_{ch_{ig}}}\Big(\Delta P_{m_{ig}} + \frac{1}{S_{ig}}\Delta\omega_i - \phi_{ig}\Delta P_{agc_i}\Big), \\ \Delta\dot{P}_{tie_{ij}} = T_{ij}\big(\Delta\omega_i - \Delta\omega_j\big), \\ ACE_i = \beta_i\Delta\omega_i + \sum_{j\in\mathcal{M}_i}\Delta P_{tie_{ij}}, \\ \Delta\dot{P}_{agc_i} = -K_{I_i}ACE_i, \end{cases} \tag{17}$$

where H_i is the equivalent inertia constant of Area i, D_i is the damping coefficient, ΔP_{m_i} and ΔP_{tie_i} are the total generated power in Area i and the total tie-line power exchanges from Area i, and ΔP_{d_i} denotes load disturbance. The term G_i is the number of participated generators in Area i, and $\mathcal{M}_i$ is the set of areas that connect to Area i. $T_{ch_{ig}}$ is the governor-turbine's time constant, S_{ig} is the droop coefficient, and T_{ij} is the synchronizing parameter between Area i and j.

In the AGC loop, note that ΔP_{agc_i} in (17) is the signal from the AGC controller for the participated generators to track the load changes, and $\phi_{i,g}$ is the participating factor, i.e., $\sum_{g=1}^{G_i}\phi_{i,g} = 1$. After receiving the frequency and tie-line power measurements, the ACE signal is computed for an integral action where β_i is the frequency bias and K_{I_i} represents the integral gain. Based on the equations in (17), the linear model of Area i can be presented as

$$\dot{\tilde{\boldsymbol{x}}}_i(t) = \boldsymbol{A}_{c,ii}\tilde{\boldsymbol{x}}_i(t) + \boldsymbol{B}_{d,i}\boldsymbol{d}_i(t) + \sum_{j\in\mathcal{M}_i}\boldsymbol{A}_{c,ij}\tilde{\boldsymbol{x}}_j(t), \tag{18a}$$

$$\boldsymbol{y}_i(t) = \boldsymbol{C}_i\tilde{\boldsymbol{x}}_i(t), \tag{18b}$$

where $\tilde{\boldsymbol{x}}_i := [\{\Delta P_{tie_{ij}}\}_{j\in\mathcal{M}_i}, \Delta\omega_i, \{\Delta P_{m_{ig}}\}_{1:G_i}, \Delta P_{agc_i}]^\top$ is the state vector that consists of area frequency, generator output, tie-line power exchange and AGC control signal of the close-loop system; $\boldsymbol{d}_i := [\Delta P_{d_i}]^\top$ denotes load disturbances. $\boldsymbol{A}_{c,ii}$ is the system matrix of Area i, $\boldsymbol{A}_{c,ij}$ is a matrix whose only non-zero element is $-T_{ij}$ in row 1 or 2 and column 3, $\boldsymbol{B}_{d,i}$ is the matrix for load disturbances. We can take an output model with high redundancy that the measurements of frequency, tie-line power exchanges, generator outputs, and AGC control signals are all measured. Then $\boldsymbol{y}_i$ is the output of Area i and $\boldsymbol{C}_i$ is the output matrix with full column rank. As noted earlier, vulnerabilities within the communication channels for frequencies and power exchanges as parts of the ACE signal may allow FDI attacks. For instance, if an attack manipulates one of tie-line power exchanges from Area i, say f_{tie_i}, then the ACE signal in (17) would be corrupted into

$$ACE_i = \beta_i\Delta\omega_i + \left(\sum_{j\in\mathcal{M}_i}\Delta P_{tie_{ij}} + f_{tie_i}\right), \tag{19}$$

which implies that FDI corruptions would affect the dynamics of controllers and consequently the involved physical system. In the end, using the state/output equation of each area, the continuous-time model of a multi-area AGC system under FDI attacks can be described in the form of (2).

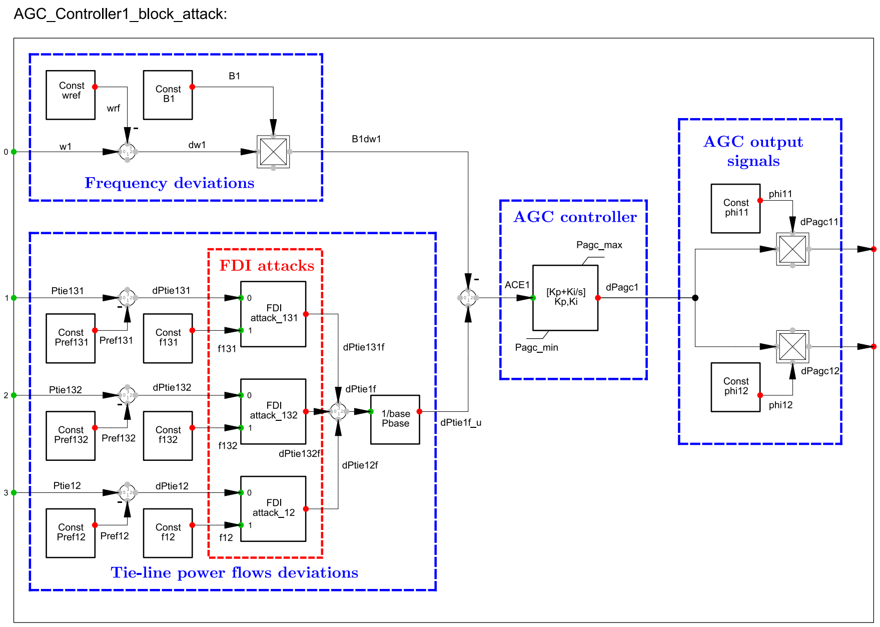


Fig. 12. The block definition of AGC in PowerFactory.

B. AGC System and FDI Attack Modeling in PowerFactory

In a high-fidelity simulator like PowerFactory, a detailed simulation of power system dynamics can be carried out that all the details of generators (e.g., synchronous machine, exciter, voltage regulator, turbine-governor unit), electric network and also controllers such as AGC can be included. Figure 1 depicts the three-area IEEE 39-bus system with AGC functions in the PowerFactory simulator. In the simulations, the dynamic generator model can consist of a synchronous machine, along with voltage regulator for the excitation system in the type of IEEE Type 1, and turbine-governor model in the type of IEEE Type G1 (steam turbine) or IEEE Type G3 (hydro turbine).

The generators in red diagrams of Figure 1 are also participating AGC control. The AGC control loop can be developed by PowerFactory's own modeling language - *DIgSILENT Simulation Language (DSL)*. For instance, Figure 12 illustrates the *block definition* of AGC in Area 1, which has several sub-blocks that collect frequencies and power exchanges between areas as the control inputs and perform AGC function to calculate signals for power settings of the participated generators in Area 1. Moreover, we build another block definition for the FDI attack model on the parts of the ACE signal. We have built the high-fidelity simulation model in the simulator PowerFactory for the 39-bus system equipped with AGC functions. The *composite frame* of AGC builds the connections between the inputs and outputs of the AGC model elements. Then the AGC *block definitions* for all areas can be created. For Figure 12 of block definition of AGC in Area 1, the four sub-blocks consist of

- *frequency deviations* block where the frequency deviations in p.u. multiplied by a bias factor are calculated;
- *tie-line power flows deviations* block which computes the tie-line power flow deviations (normalized in p.u.) on the side of Area 1 for the power part of ACE;
- *AGC controller* block which performs the ACE calculation and the integral action to generate the tuning signal (ΔP_{agc_i}) for power settings of participated generators. The saturation effects are considered that the limits of P_{agc}^{min} and P_{agc}^{max} are added for the tuning signal;
- *AGC output signals* block where the tuning signals for the participated generators in Area 1 for AGC are calculated based on each generator's participating factor.

The above block definitions are modeled using the *Standard Macros* of PowerFactory's global *Library*. Moreover, in Figure 12, another block definition (in red diagram) corresponds to the FDI attack model for the study of this article,

- *FDI attacks* block where the FDI attack is implemented. Each block captures the feature of the stationary FDI attack, i.e., the attack occurs as a constant bias injection ($f[k] = f$) on measurements at a specific time step, and it remains unchanged since then. This block can add an "false" injection into the existing signal. One can specify the occurrence time and the attack values. This block definition is achieved by using the *digexfun* interface. With *digexfun*, we can define a specific DSL function (in C++) and create a dynamic link library *digexfun_*.dll* that PowerFactory can load.

References

[1] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[2] X. Luo, Q. Yao, X. Wang, and X. Guan, "Observer-based cyber attack detection and isolation in smart grids," *Int. J. Electr. Power Energy Syst.*, vol. 101, pp. 127–138, Oct. 2018.

[3] M. Ashrafuzzaman, S. Das, Y. Chakhchoukh, S. G. Shiva, and F. T. Sheldon, "Detecting stealthy false data injection attacks in the smart grid using ensemble-based machine learning," *Comput. Security*, vol. 97, Oct. 2020, Art. no. 101994,

[4] T. Wei, X. Chen, X. Li, and Q. Zhu, "Model-based and data-driven approaches for building automation and control," in *Proc. IEEE/ACM ICCAD*, 2018, pp. 1–8.

[5] K. Pan, P. Palensky, and P. M. Esfahani, "From static to dynamic anomaly detection with application to power system cyber security," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1584–1596, Mar. 2020.

[6] P. Sauer and M. Pai, *Power System Dynamics and Stability*. Hoboken, NJ, USA: Prentice Hall, 1998.

[7] S. Li, Y. Yılmaz, and X. Wang, "Quickest detection of false data injection attack in wide-area smart grids," *IEEE Trans. Smart Grid*, vol. 6, no. 6, pp. 2725–2735, Nov. 2015.

[8] J. Zhao, L. Mili, and M. Wang, "A generalized false data injection attacks against power system nonlinear state estimator and countermeasures," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4868–4877, Sep. 2018.

[9] J. Tian, R. Tan, X. Guan, Z. Xu, and T. Liu, "Moving target defense approach to detecting stuxnet-like attacks," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 291–300, Jan. 2020.

[10] S. Lakshminarayana, E. V. Belmega, and H. V. Poor, "Moving-target defense against cyber-physical attacks in power grids via game theory," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5244–5257, Nov. 2021.

[11] M. Nyberg and E. Frisk, "Residual generation for fault diagnosis of systems described by linear differential-algebraic equations," *IEEE Trans. Autom. Control*, vol. 51, no. 12, pp. 1995–2000, Dec. 2006.

[12] A. F. Taha, J. Qi, J. Wang, and J. H. Panchal, "Risk mitigation for dynamic state estimation against cyber attacks and unknown inputs," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 886–899, Mar. 2018.

[13] A. Ameli, A. Hooshyar, E. F. El-Saadany, and A. M. Youssef, "Attack detection and identification for automatic generation control systems," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4760–4774, Sep. 2018.

[14] M. Khalaf, A. Youssef, and E. El-Saadany, "Joint detection and mitigation of false data injection attacks in AGC systems," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 4985–4995, Sep. 2019.

[15] J. Qi, A. F. Taha, and J. Wang, "Comparing Kalman filters and observers for power system dynamic state estimation with model uncertainty and malicious cyber attacks," *IEEE Access*, vol. 6, pp. 77155–77168, 2018.

[16] S. X. Ding, *Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools*. Heidelberg, Germany: Springer, 2008.

[17] P. M. Esfahani and J. Lygeros, "A tractable fault detection and isolation approach for nonlinear systems with probabilistic performance," *IEEE Trans. Autom. Control*, vol. 61, no. 3, pp. 633–647, Mar. 2016.

[18] P. Palensky, A. van der Meer, C. Lopez, A. Joseph, and K. Pan, "Applied cosimulation of intelligent power systems: Implementing hybrid simulators for complex power systems," *IEEE Ind. Electron. Mag.*, vol. 11, no. 2, pp. 6–21, Jun. 2017.

[19] J. J. Q. Yu, Y. Hou, and V. O. K. Li, "Online false data injection attack detection with wavelet transform and deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3271–3280, Jul. 2018.

[20] A. Ayad, H. E. Z. Farag, A. Youssef, and E. F. El-Saadany, "Detection of false data injection attacks in smart grids using recurrent neural networks," in *Proc. IEEE PES ISGT Conf.*, 2018, pp. 1–5.

[21] Y. Wadhawan, A. AlMajali, and C. Neuman, "A comprehensive analysis of smart grid systems against cyber-physical attacks," *Electronics*, vol. 7, no. 10, p. 249, 2018.

[22] J. Sakhnini, H. Karimipour, and A. Dehghantanha, "Smart grid cyber attacks detection using supervised learning and heuristic feature selection," in *Proc. SEGE*, 2019, pp. 108–112.

[23] J. Hao, R. J. Piechocki, D. Kaleshi, W. H. Chin, and Z. Fan, "Sparse malicious false data injection attacks and defense mechanisms in smart grids," *IEEE Trans. Ind. Informat.*, vol. 11, no. 5, pp. 1–12, Oct. 2015.

[24] C. Wang, S. Tindemans, K. Pan, and P. Palensky, "Detection of false data injection attacks using the autoencoder approach," in *Proc. Int. Conf. Probabilistic Methods Appl. Power Syst. (PMAPS)*, Mar. 2020, pp. 1–6.

[25] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 8, pp. 1773–1786, Aug. 2016.

[26] S. A. Foroutan and F. R. Salmasi, "Detection of false data injection attacks against state estimation in smart grids based on a mixture gaussian distribution learning method," *IET Cyber-Phys. Syst. Theory Appl.*, vol. 2, no. 4, pp. 161–171, 2017.

[27] Y. Zhang, J. Wang, and B. Chen, "Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 623–634, Jan. 2021.

[28] M. N. Kurt, O. Ogundijo, C. Li, and X. Wang, "Online cyber-attack detection in smart grid: A reinforcement learning approach," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5174–5185, Sep. 2019.

[29] K. Tidriri, N. Chatti, S. Verron, and T. Tiplica, "Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges," *Annu. Rev. Control*, vol. 42, pp. 63–81, Sep. 2016.

[30] J. Zhang and A. Domínguez-García, "On the impact of communication delays on power system automatic generation control performance," in *Proc. NAPS*, 2014, pp. 1–6.

[31] P. Kundur, N. Balu, and M. Lauby, *Power System Stability and Control*. New York, NY, USA: McGraw-Hill Educ., 1994. [Online]. Available: https://books.google.nl/books?id=2cbvyf8Ly4AC

[32] E. Rakhshani, D. Remon, A. M. Cantarellas, J. M. Garcia, and P. Rodriguez, "Virtual synchronous power strategy for multiple HVDC interconnections of multi-area AGC power systems," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 1665–1677, May 2017.

[33] C. W. Ten, C. C. Liu, and G. Manimaran, "Vulnerability assessment of cybersecurity for SCADA systems," *IEEE Trans. Power Syst.*, vol. 23, no. 4, pp. 1836–1846, Nov. 2008.

[34] K. Pan, A. Teixeira, C. D. López, and P. Palensky, "Co-simulation for cyber security analysis: Data attacks against energy management system," in *Proc. IEEE SmartGridComm*, 2017, pp. 253–258.

[35] C. Chen, K. Zhang, K. Yuan, L. Zhu, and M. Qian, "Novel detection scheme design considering cyber attacks on load frequency control," *IEEE Trans. Ind. Informat.*, vol. 14, no. 5, pp. 1932–1941, May 2018.

[36] 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.

[37] K. Ogata, *Discrete-Time Control Systems*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1995.

[38] S. A. Vavasis, *Complexity Theory: Quadratic Programming*. Boston, MA, USA: Springer, 2001, pp. 304–307. [Online]. Available: https://doi.org/10.1007/0-306-48332-7_65

[39] H. Bevrani, *Robust Power System Frequency Control* (Power Electronics and Power Systems). Boston, MA, USA: Springer, 2008.

[40] "PowerFactory: 39 bus New England system," DIgSILENT GmbH, Gomaringen, Germany, Rep. r1338, 2018.

[41] P. M. Esfahani, T. Sutter, and J. Lygeros, "Performance bounds for the scenario approach and an extension to a class of non-convex programs," *IEEE Trans. Autom. Control*, vol. 60, no. 1, pp. 46–58, Jan. 2015.

[42] L. Schenato, "To zero or to hold control inputs with lossy links?" *IEEE Trans. Autom. Control*, vol. 54, no. 5, pp. 1093–1099, May 2009.

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

[44] A. Hoehn and P. Zhang, "Detection of replay attacks in cyber-physical systems," in *Proc. Amer. Control Conf.*, 2016, pp. 290–295.

Kaikai Pan (Member, IEEE) received the B.Sc. and M.Sc. degrees in information engineering, measuring, and control from Beihang University, Beijing, China, in 2012 and 2015, respectively, and the Ph.D. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2020. He is currently an Assistant Professor with the College of Electrical Engineering, Zhejiang University, Hangzhou, China. He was a Research Fellow with the Electrical Sustainable Energy department, Delft University of Technology. His research interests include attack or anomaly detection, cyber risk assessment, power industrial control systems security, Internet-of-Things security, cyber-physical systems modeling, and co-simulation techniques. He is active in committees, such as IEEE as TPC members and a reviewer for various journals and conferences.

Peter Palensky (Senior Member, IEEE) was born in Austria in 1972. He received the M.Sc. degree in electrical engineering and the Ph.D. degree from the Vienna University of Technology, Austria, in 1997 and 2001, respectively. He co-founded an Envidatec, a German startup on energy management and analytics, and joined the Lawrence Berkeley National Laboratory, Berkeley, CA, USA, as a Researcher, and the University of Pretoria, South Africa, in 2008. In 2009, he became the Head of Business Unit on sustainable building technologies with the Austrian Institute of Technology (AIT), and later the first Principle Scientist for complex energy systems. In 2014, he was appointed as a Full Professor of Intelligent Electric Power Grids with TU Delft. His main research fields are energy automation networks, smart grids, and modeling intelligent energy systems. He is active in international committees, such as ISO or CEN and serves as an IEEE IES AdCom Member-at-Large in various functions for the IEEE. He is the Editor-in-Chief for the *IEEE Industrial Electronics Magazine*, an associate editor for several other IEEE publications, and regularly organizes IEEE conferences.

Peyman Mohajerin Esfahani received the B.Sc. and M.Sc. degrees from the Sharif University of Technology, Iran, and the Ph.D. degree from ETH Zurich. He is currently an Assistant Professor with the Delft Center for Systems and Control, Delft University of Technology. Prior to joining TU Delft, he held several research appointments with EPFL, ETH Zurich, and MIT from 2014 to 2016. His research interests include theoretical and practical aspects of decision-making problems in uncertain and dynamic environments, with applications to control and security of large-scale and distributed systems. He was one of the three finalists for the Young Researcher Prize in Continuous Optimization awarded by the Mathematical Optimization Society in 2016, and was a recipient of the 2016 George S. Axelby Outstanding Paper Award from the IEEE Control Systems Society.