RADM: A Risk-Aware DER Management Framework with Real-time DER Trustworthiness Evaluation

Yaodan Hu
University of Florida
cindy.hu@ufl.edu

Xiaochen Xian
University of Florida
xxian@ufl.edu

Yier Jin
University of Florida
yier.jin@ece.ufl.edu

ABSTRACT

The increasing penetration level of distributed energy resources (DERs) substantially expands the attack surface of the modern power grid. By compromising DERs, adversaries are capable of destabilizing the grid and potentially causing large-area blackouts. Due to the limited administrative control over DERs, constrained computational capabilities, and possible physical accesses to DERs, current device level defenses are insufficient to defend against malicious attacks on DERs. To compensate the shortcomings of device level defenses, in this paper, we develop a system-level risk-aware DER management framework (RADM) to mitigate the attack impacts. We propose a metric, *trust score*, to dynamically evaluate the trustworthiness of DERs. The *trust scores* are initialized with offline trust scores derived from static information and then regularly updated with online trust scores derived from a physics-guided Gaussian Process Regressor using real-time data. The *trust scores* are integrated into the grid control decision making process by balancing the grid performance and the security risks. Extensive simulations are conducted to justify the effectiveness of the proposed method.

CCS CONCEPTS

• **Security and privacy → Trust frameworks**.

KEYWORDS

DER Security, Gaussian Process Regression, Smart Grid, Cyber-Physical System, Trustworthiness Evaluation, Trust Score

ACM Reference Format:
Yaodan Hu, Xiaochen Xian, and Yier Jin. 2021. RADM: A Risk-Aware DER Management Framework with Real-time DER Trustworthiness Evaluation. In *ACM/IEEE 12th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2021) (ICCPS '21), May 19–21, 2021, Nashville, TN, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3450267.3450536

1 INTRODUCTION

The traditional power grid is undergoing a massive change through the integration of distributed energy resources (DERs) [9]. DERs represent the power generation devices or controllable loads spreading in the distribution system, such as renewable energy harvesting devices like photovoltaic (PV) systems, electrical vehicles, and electricity storage, etc. In 2018, the capacity of renewable energy, which constitutes the majority of DERs, has reached 20.5% of the total electricity generating capacity in the U.S. [14]. With the significant increase in the DER penetration level, DERs have been taking an active role in grid control operations such as demand response and frequency stabilization, due to its communication and control capabilities similar to many other Internet of Things (IoT) devices.

Although DERs enrich the power grid with increasing autonomy and flexibility, they at the same time lead to a substantially expanded attack surface of the power system [7], which may cause the uneconomical dispatch of power, unstable grid status, or even large-area blackouts [4, 5, 23]. The threats induced by the integration of DERs are mainly because of two reasons. First, like many other IoT devices, most DERs are computationally constrained and lack of security in design. Thus the DERs are implemented with no or poor security defenses. Second, the connectivity between DERs and the grid makes more types and larger scales of cyber-physical attacks possible. As most DERs are owned and controlled by consumers and third-parties, they may be inappropriately operated, leaving vulnerabilities for attackers to exploit.

The asymmetry between the importance and the reliability of DERs makes DERs attractive to attackers [4, 7]. To protect DERs against malicious attacks, various approaches have been developed in the literature, such as the usage of trusted execution environments (TEEs) [21], voltage-state of charge feature [30], cryptography [15], sliding mode control [12], etc. Nevertheless, there is still a lack of adequate protections on DERs due to the following reasons. First of all, defenses relying on hardware modifications [21, 30] are ineffective as the number of DERs is enormous; 2 million residential PV systems have been installed in the U.S. at the end of 2018. Considering a 20-year life span of solar panels, the poor scalability limits the application of protections using hardware modifications. Moreover, cryptography is not an ideal solution due to the limited computational capability. To avoid long latency, only naive cryptography mechanisms can be implemented on DER devices, e.g., the simplest 128-bit Advanced Encryption Standard (AES) with a mean latency of 4.05ms [15]. The latency for more complicated mechanisms is expected to be much longer. Nonetheless, systems such as substations require latency on the order of 10ms, which can hardly be satisfied if more sophisticated cryptography is required or grid support functions are implemented. In the literature, there are also defense methodologies which modify the control strategies embedded in the DER firmware [12]. Such methods require less computational resources and can be implemented through wireless firmware update. However, since most DERs are installed outside, the attacker may gain physical access to the device [4], and thus,

ICCPS '21, May 19–21, 2021, Nashville, TN, USA

ACM ISBN 978-1-4503-8353-0/21/05…$15.00
https://doi.org/10.1145/3450267.3450536

methods relying on the control strategies at the DER side might not be effective.

Due to the limited administrative controls over DERs, constrained computational capabilities, and the massive volume of DERs, implementing defenses only on DERs are insufficient against malicious attacks. To overcome the limitations of the current defense methods, we introduce an additional system defense layer on top of the device level defenses. Since the solar energy is among the most widely used DERs, in this paper, we use PV systems as a case study of DERs. We consider an attacker capable of compromising the data integrity of DERs and propose a framework that protects the operations of the grid in the presence of such attacks. Specifically,

- We propose RADM, a risk-aware DER management framework, to robustly integrate DERs into the power grid and increase the resilience of the grid. We use the *trust score* to quantify the probability of attacks on a DER, i.e., the trustworthiness of a DER, and leverage the *trust scores* for real-time risk-aware DER management. By doing so, the grid is capable of maintaining normal operations even in the appearance of attacks.
- Since the attack launch time is not determined, the *trust scores* are dynamically estimated with a Bayesian framework. The framework initializes the *trust scores* with offline trust scores estimated from static information, such as DER firmware and grid topology, to present a general assessment of the DER security levels. The framework then updates the *trust scores* with online trust scores utilizing real-time information, such as weather information and solar power generations, for a timely understanding of the DER status.
- To update the *trust scores* and obtain the real-time trustworthiness of DERs, we propose a physics-guided Gaussian Process Regressor integrating physical domain knowledge and data-driven patterns. By leveraging the DER physical model, the regressor prediction results are regulated by physical laws and thus responsive to attacks. The data-driven patterns learned from historical observations allow to enhance the prediction power by cross-checking nearby DERs that share similar generation patterns.
- Simulations are conducted to justify the effectiveness of RADM. The results prove that our method can mitigate the attack impacts with slight performance degradation.

The rest of the paper is organized as follows: Section 2 discusses the current literature of the DER security and Section 3 introduces the background knowledge of the paper. In Section 4, we describe the system and the threat model considered, and formulate the problem. The detailed trustworthiness evaluation framework is introduced in Section 5. In Section 6, we use voltage regulation as a case study and show how the DER trustworthiness is integrated into the control decisions. Section 7 presents the performance of the proposed work and in Section 8, we conclude our work.

2 RELATED WORK

There have been several standards and guidelines addressing the security of power systems, such as the NIST Framework for Improving Critical Infrastructure Cybersecurity [8] and IEEE C37.240 [1]. However, since power systems such as substations are within administrative controls and have sufficient computational resources, only few standards addressed the unique challenges of DERs. The IEEE 1547 Standards [2] is designed specifically for DERs but no practice for DER security has been recommended yet.

To defend against attacks on DERs, various methods have been proposed. These methods can be divided into two categories: hardware-based and software-based. The hardware-based defenses leverage hardware components to be the root-of-trust. In [21], Sebastian et al. proposed to utilize the hardware-enforced trusted execution environments (TEEs) and utilized the secure storage and cryptographic functions of TEEs to reduce the attack surface and guarantee the data integrity of DERs. Besides, Zografopoulos et al. [30] explored the correlations between the state-of-charge (SoC) and the voltage measurements from DER battery energy storage system (BESS) and designed a DER authentication mechanism based on the challenge-reply sequences (CRSeqs). There have also been various works attempting to develop cryptographic modules. Lai et al. built a cryptographic module with keys embedded in the trusted platform modules (TPM) and examined the feasibility of implementing different cryptography methods on DERs [15]. Hupp et al. [13] developed Module-OT which was integrated in the transport layer of the Open Systems Interconnection (OSI) model. An average latency of 4 ms was observed in [15] and 6ms in [13], which barely fulfilled the latency requirement of substations. Since the hardware-based defenses require modifications to the DER hardware, not only the scalability is poor but also the cost is prohibitively high because of the large population and the long life span of DERs.

Compared with the hardware-based defenses, the software-based defenses are more affordable because the software can be updated remotely. Cryptography has also been suggested in the software-based defenses [9, 20]. As an emerging technology, blockchains have also been applied to DERs [19]. Nevertheless, similar to the cryptographic modules in the hardware-based defenses, concerns about potential latency still exist in these methods due to the limited computational resources of DERs, especially for blockchains, which have heavy computational overheads. Besides cryptography, different control strategies have also been developed. Gholami et al. designed a sliding mode observer to estimate the attack vector and compensate the manipulated data [12]. Thus the observer is capable of forcing the safe operation of DERs. Few papers considered the grid resilience under attacks, i.e., how to maintain the operation of the grid in case of successful attacks. Srikantha et al. in [24] constructed the control policies by formulating a two-player zero-sum differential game between the control center and the attacker. The authors demonstrated that the attack impacts could be mitigated as long as a set of uncompromised components exist. However, the control center could only win the game and maintain the grid stability when the grid scale is smaller than 39 buses.

3 BACKGROUND

In this section, we introduce the background knowledge required in the trustworthiness evaluation framework. We leverage Gaussian Process Regression (GPR) to validate the trustworthiness of the DER measurements, in our case, the PV generation reports. To mitigate the impacts of spoofed reports and enhance the prediction

accuracy, we combine the DER physical model, in our case, the PV array circuit model, with GPR.

3.1 Gaussian Process Regression

A Gaussian process (GP) is a stochastic process in which any finite collection of the random variables in the process follows a multivariate normal distribution. As a powerful regression algorithm, GPR is able to learn an unknown function and also give a reliable estimate of their own uncertainty, with a standard function format $y = f(x) + \epsilon$ where ϵ is the independent noise following the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. In GPR, $f(x)$ is assumed to be a random variable from a GP. That is, for any finite collection $\mathbf{x}$, $f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K)$, where $m(\mathbf{x}) = E[f(\mathbf{x})]$. K is the covariance matrix and $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$, where $k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$ is the kernel function measuring the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ with hyper-parameter $\boldsymbol{\theta}$.

Given a training dataset $\{\mathbf{x}, \mathbf{y}\}$ and new observations $\mathbf{x}_*$, the objective of GPR is to estimate the posterior distribution $p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}_*)$. Based on the definition of GP, the joint distribution of $\mathbf{y}$ and $\mathbf{y}_*$ given $\mathbf{x}$ and $\mathbf{x}_*$ is given as:

$$\begin{aligned} p(\mathbf{y}; \mathbf{y}_*|\mathbf{x}, \mathbf{x}_*) &\sim \mathcal{N}\left(m(\mathbf{x}), \begin{pmatrix} K + \sigma^2 I & K_* \\ K_*^T & K_{**} \end{pmatrix}\right), \\ K_* &= k(\mathbf{x}, \mathbf{x}_*; \boldsymbol{\theta}), \\ K_{**} &= k(\mathbf{x}_*, \mathbf{x}_*; \boldsymbol{\theta}). \end{aligned} \tag{1}$$

Therefore, $p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}_*) \sim \mathcal{N}(\mu, \Sigma)$, in which

$$\begin{aligned} \mu &= E[f(\mathbf{x})] = m(\mathbf{x}) + K_*^T (K + \sigma^2 I)^{-1} \mathbf{y}, \\ \Sigma &= V[f(\mathbf{x})] = K_{**} - K_*^T (K + \sigma^2 I)^{-1} K_*. \end{aligned} \tag{2}$$

3.2 Single-Diode PV Array Circuit Model

PV systems are composed of inverters and PV panels. The PV panels can be regarded as a series of solar cells. According to [26], given a PV array with N_P strings in parallel and N_S cells in series, the output current

$$I = N_P I_{irr} - N_P I_0 \left[\exp\left(\frac{q\left(V + I\frac{N_S}{N_P} R_S\right)}{N_S n \kappa T} \right) - 1 \right] - \frac{V + I\frac{N_S}{N_P} R_S}{\frac{N_S}{N_P} R_P}, \tag{3}$$

in which I_{irr} is the photocurrent, V is the output voltage, I_0 is the diode saturation current, R_S is the series resistance, and T is the cell temperature. $q = 1.602 \times 10^{-19}$C is the electronic charge, and $\kappa = 1.3806503 \times 10^{-23}$J/K is the Boltzmann's constant. n is the diode ideality factor and almost remains constant w.r.t. operation conditions. The derivation of the equation is elaborated in [26].

Typically each PV array is equipped with a Maximum Power Point Tracking (MPPT) controller and operated at the maximum power P. Here we consider the incremental conductance algorithm because of its efficient and stable tracking performance. Thus, when the maximum power is reached, we have

$$\frac{dP}{dV} = \frac{d(IV)}{dV} = I + V\frac{dI}{dV} = 0.$$

That is,

$$\frac{I}{V} = \frac{\frac{qN_P I_0}{N_S n \kappa T} \exp\left(\frac{q\left(V + I\frac{N_S}{N_P} R_S\right)}{N_S n \kappa T} \right) + \frac{1}{\frac{N_S}{N_P} R_P}}{1 + \frac{qI_0 R_S}{n\kappa T} \exp\left(\frac{q\left(V + I\frac{N_S}{N_P} R_S\right)}{N_S n \kappa T} \right) + \frac{R_S}{R_P}}. \tag{4}$$

By solving Eq. 3 and Eq. 4, I and V can be determined, and the PV generation power $P = IV$.

4 PROBLEM DESCRIPTION

In this section, we formulate the problem and present a summary in Figure 1. Specifically, we introduce the system model and the threat model we tangle with and present a brief introduction to the proposed DER management framework.

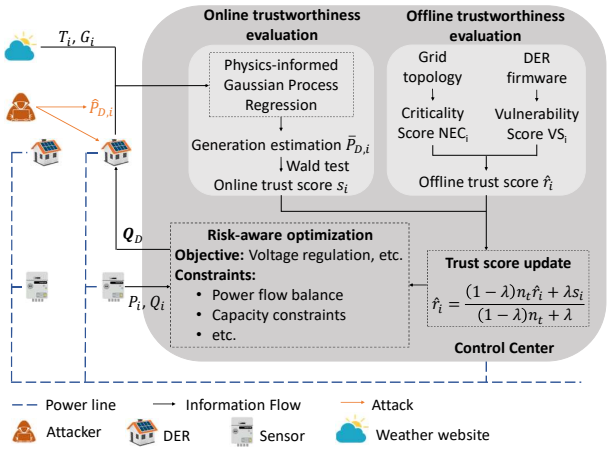

Figure 1: The architecture of RADM. T_i and G_i are the weather information. $\hat{P}_{D,i}$ is the reported DER measurements and in this case, the reported power generations from PV systems. P_i and Q_i are the sensor measurements such as active/reactive power flows. Q_D is the request to DERs.

4.1 System Model

In this paper, we consider a simplified grid system in which DERs are coordinated by a control center such as an independent system operator (ISO). Without loss of reality, we assume that DERs are required to report their manufacturing models, sizes (i.e., N_P and N_S), and the locations before installation. Since the data sheets of different DER manufacturing models are open to public, the control center knows the operation status at the Standard Reference Conditions (SRC). We also assume that the control center has full knowledge of the grid, such as the grid topology and the line admittance. We represent the grid as a weighted graph with each bus (such as a substation) being a node and the nodal admittance matrix $\tilde{\mathbf{Y}} \in \mathbb{C}^{n_b \times n_b}$ being the weights. In the rest of the paper, we use "bus" and "node" interchangeably. $\tilde{\mathbf{Y}}_{ik}$ denotes the admittance between bus i and k, and n_b denotes the number of buses in the grid. If there is no physical power line connecting buses i and k, $\tilde{\mathbf{Y}}_{ik} = 0$. $\tilde{\mathbf{Y}}$ not only represents the physical connection of the grid, but also indicates the power transfer capability of each power line. With the settings above, we have $\tilde{\mathbf{Y}}\tilde{\mathbf{V}} = \tilde{\mathbf{I}}$, in which $\tilde{\mathbf{V}}, \tilde{\mathbf{I}} \in \mathbb{C}^{n_b \times 1}$ are the

vectors of the voltages and the currents at each bus, respectively. We also assume the power system is balanced and thus we have:

$$\tilde{\mathbf{S}}_i = \tilde{\mathbf{V}}_i \sum_{k=1}^{n_b} (\tilde{\mathbf{Y}}_{ik} \tilde{\mathbf{V}}_k)^*, \tag{5}$$

where $\tilde{\mathbf{S}}_i$ denotes the power injection at bus i.

We assume that two-way communication links are established between the control center and DERs, through either wired or wireless communications. There are also a set of sensors, such as smart meters and Phasor Measurement Units (PMUs), installed across the grid for monitoring purposes. During operations, DERs and sensors report their status and measurements, e.g., the power generations, active/reactive power flow injections, voltage magnitudes/angles, etc., to the control center. The control center also has access to the weather data, including the solar irradiation and the temperature data from the meteorological authority or public websites. In order to maintain the grid stably and economically, the control center can issue tasks, such as active/reactive power injections, to DERs, according to the measurements from sensors and DERs.

4.2 Threat Model

In the threat model, we assume the adversary conducts attacks on DERs. Compared with the control center, DERs are prone to be attacked because the control center is much more powerful in computing than DERs and can have various advanced security mechanisms implemented. To mislead the control center, the adversary compromises the data integrity of the DERs measurements by exploiting vulnerabilities in DERs or the communication protocols. With the tampered DER measurements, the grid will be operated with inaccurate control decisions, which may lead to economic losses or even blackouts.

We assume the adversary can compromise the DERs by physical attacks [4], tampering DER firmware and/or spoofing the DER measurements packets [7]. Due to limited resources, the adversary can compromise at most k DERs. Thus the adversary has to deliberately identify the target DERs. The attacker is more willing to attack the DERs with poor security defenses while located at critical buses. Since different DER manufacturing models have different implementations, some DERs are more likely to be penetrated because of the potential vulnerabilities in their firmware. Besides, to maximize the attack impacts, the adversary would attack the DERs located at the critical buses of the grid. Moreover, the attacker intends to pick victim DERs in an area instead of spreading over a large area. By doing so, the adversary can conduct coordinated attacks like that in [23], which has the potential to cause large-area blackouts.

4.3 Overview of RADM

Considering the possible attacks on the DERs, we intend to maintain the normal operations of the grid even in the presence of attacks. To achieve the goal, we develop a framework for robust DER management, RADM, as depicted in Figure 1. In the framework, we use *trust scores* to dynamically quantify the trustworthiness of each DER and operate the grid based on the DER *trust scores*. The *trust scores* indicate the probability that a DER is *not* under attack and the estimation of *trust scores* is summarized in Algorithm 1. By leveraging the *trust scores*, the control center is facilitated with more precise awareness of the DER security.

Algorithm 1 Trust Score Estimation

Input: DER firmware, $\mathbf{S}$, $\tilde{\mathbf{Y}}$, n_t, DER datasheets and locations, historical weather information, historical DER measurements, real-time weather information $\{G_i(t)\}$, $\{T_i(t)\}$, real-time DER measurements $\{\hat{P}_{D,i}(t)\}$

Output: $\{\hat{r}_i\}$

1: **Initialize:**
2: $\{\alpha_i\}$, $\{\beta_i\}$, $\{\hat{r}_i\}$ = OfflinePhase(DER firmware, $\mathbf{S}$, $\tilde{\mathbf{Y}}$, n_t);
3: $\{\theta_i(t)\}$, $\{\bar{K}_i(t)\}$, $\{K_i^{PZ}(t)\}$, $\{K_i^{ZZ}(t)\}$ = GPRTraining(DER datasheets and locations, historical DER measurements);
4:
5: **repeat**
6: $\{s_i(t)\}$ = OnlinePhase($\{G_i(t)\}$, $\{T_i(t)\}$, $\{\hat{P}_{D,i}(t)\}$, $\{\theta_i(t)\}$, $\{\bar{K}_i(t)\}$, $\{K_i^{PZ}(t)\}$, $\{K_i^{ZZ}(t)\}$);
7: **Update** $\{\alpha_i\}$ and $\{\beta_i\}$ according to Eq. 22 and $\{\hat{r}_i\}$ according to Eq. 23.
8: **return** $\{\hat{r}_i\}$;
9: **until** No new readings come in

Denote the set of DERs as $\mathbf{N} = \{1, \ldots, n\}$ where n is the number of DERs. We denote the *trust score* of DER i as r_i and model the status of each DER, i.e., attacked or not, at each time, as a Bernoulli trial with the parameter r_i. Since the start time of the attack is unknown, the control center dynamically update their belief on r_i based on a Bayesian framework. Compared with other trust evaluation frameworks such as the Dempster–Shafer theory [27], the Bayesian framework is easier to implement. The Bayesian framework is composed of an offline phase and an online phase. The offline phase provides us with a general belief on the likelihood a DER might be attacked, and the online phase tests the real-time DER status based on dynamic data inputs. The former enhances the robustness of the trustworthiness evaluation system, while the latter can provides an up-to-date understanding of DERs.

Without loss of generality, we assume the *trust score* of DER i follows a conjugate prior distribution, i.e., $r_i \sim \text{Beta}(\alpha_i, \beta_i)$. The hyper-parameters α_i and β_i, and the *trust score* are initially estimated with the offline trust score $\hat{r}_i$ in the offline phase (Algorithm 1, line 2). The offline trust score is derived from the DER vulnerability score VS_i and the DER criticality score NEC_i. VS_i reflects the adversary's capability of manipulating the DER and is calculated according to the DER firmware analysis results. NEC_i reflects the potential damages on the grid that an adversary can cause and is calculated based on the grid topology. Since the start of an attack is unpredictable, α_i and β_i are updated dynamically with the online trust score s_i to obtain a timely estimation of the *trust score* $\hat{r}_i$ (Algorithm 1, line 6-7). The online trust score s_i is the test result from a physics-guided Gaussian Process Regressor, which combines the physics-based model with a data-driven Gaussian Process (GP). Most computations of the *trust score* estimation scheme is processed offline, including the offline trust estimation and the GP training. Thus, the scheme induces little computational overhead during run-time and is efficient for real-time implementations.

When generating requests for DERs regarding different operations such as voltage regulations or economic dispatches, we formulate the tasks as optimization problems with the estimated *trust*

score $\hat{r}_i$ included to balance the impacts of attacks. The details of the framework will be introduced in the following sections.

5 TRUST SCORE ESTIMATION

5.1 Offline Estimation

As described in Algorithm 2, we provide a general belief on the distribution of the *trust score* r_i by quantifying α_i and β_i of the prior distribution based on the DER vulnerability score VS_i and the DER criticality score NEC_i.

Algorithm 2 Offline Trust Score Estimation

1: **function** OfflinePhase(DER firmware, $\mathbf{S}$, $\tilde{\mathbf{Y}}$, n_t)
2: **for** $i \in \mathbf{N}$
3: Construct $\mathbf{v}^i$ through automatic firmware vulnerability detection;
4: $VS_i = \min(\frac{||\mathbf{v}^i \circ \mathbf{S}||_2}{10}, 1)$;
5: Calculate NEC_i with $\tilde{\mathbf{Y}}$, the set of generators and the set of loads;
6: $\hat{r}_i = (1 - VS_i \times NEC_i)$;
7: $\alpha_i = n_t \hat{r}_i, \quad \beta_i = n_t - \alpha_i$;
8: **end for**
9: **return** $\{\alpha_i\}, \{\beta_i\}, \{\hat{r}_i\}$;
10: **end function**

5.1.1 DER vulnerability score. Due to the heterogeneity of the devices in the grid, different devices may have different functionalities and implementations, and thus different vulnerabilities may lie in different devices. Since DERs with more severe vulnerabilities are more likely to be compromised, we check the possible vulnerabilities in DER firmware and derive the DER vulnerability score VS_i for each DER i (Algorithm 2, line 3-4). We choose the Common Vulnerability Scoring System (CVSS) to quantify the severity of a vulnerability. We denote the vector of the severity of vulnerabilities using $\mathbf{S} \in [0, 10]^{n_S}$, in which the k-th element $\mathbf{S}_k$ denotes the severity score of the vulnerability k and n_S is the number of vulnerabilities recorded. Note that the construction of $\mathbf{S}$ is static and only requires regular update, therefore, no run time overhead is induced. By automatically exploring the vulnerabilities in device firmware (existing work can be found in [16]), a boolean vector $\mathbf{v}^i \in \{0, 1\}^{n_S}$ can be constructed for DER i (Algorithm 2, line 3). $\mathbf{v}^i_k$, i.e., the k-th element in $\mathbf{v}^i$, is a boolean variable indicating whether the vulnerability k exists in DER i. To justify how likely an adversary is capable of compromising DER i, we integrate all the detected vulnerabilities by defining the DER vulnerability score VS_i as the scaled distance between the state of the DER i, $\mathbf{v}^i$, to the origin, i.e., a secure state with no vulnerability (Algorithm 2, line 4):

$$VS_i = \min\left(\frac{||\mathbf{v}^i \circ \mathbf{S}||_2}{10}, 1\right), \tag{6}$$

where $\mathbf{v}^i \circ \mathbf{S}$ is the Hadamard product, i.e., the elemental-wise product, of $\mathbf{v}^i$ and $\mathbf{S}$. $||\cdot||_2$ indicates the L_2 norm.

5.1.2 DER criticality score. DERs locate at different places will have different impacts on the grid if compromised. Therefore, besides VS_i which determines how likely an adversary can manipulate the device, the criticality of a device to grid that determines how much an adversary can impact the grid, is also an imperative component of the trustworthiness assessment of a DER. Since the more critical a node is, the more likely it will be the target of an adversary, a critical node is expected to have a relatively low *trust score*. Many metrics have been proposed to evaluate the node criticality [17, 28]. Here we adopt the Node Electrical Centrality (NEC) proposed in [17] (Algorithm 2, line 5). The NEC of node i, NEC_i, reflects the impact of node i to the grid if the node is removed. NEC_i is a weighted average of the electrical betweenness centrality and the eigenvector centrality. Instead of treating all nodes equally, the authors assign different attributes (generators and/or loads) to the nodes. With such setting, the electrical betweenness of a node is defined as the weighted sum of currents flowing through the node w.r.t each generator-load pairs. Compared with the standard definition of betweenness centrality in graph theory, the definition of the electrical betweenness centrality is more reasonable because the currents do not flow along the shortest paths only. Due to the limited space, we will not elaborate how NEC_i is calculated here.

5.1.3 Trust score in the offline phase. VS_i infers the probability that an adversary is capable of compromising a DER, and the criticality score NEC_i infers the probability that an adversary would like to attack the device. As assumed in the threat model (Section 4.2), the attacker is more likely to attack the DERs with poor security defenses while located at critical buses. Therefore, $VS_i \times NEC_i$ indicates the probability that a DER will be compromised, and the *trust score* of DER i, i.e., the probability that the DER will *not* be compromised, is computed as (Algorithm 2, line 6)

$$\hat{r}_i = 1 - VS_i \times NEC_i. \tag{7}$$

Statistically, we expect $\hat{r}_i$ to be the initial belief of the trustworthiness of DER i and can derive that (Algorithm 2, line 7)

$$\begin{aligned} \hat{r}_i = E(r_i) = \frac{\alpha_i}{\alpha_i + \beta_i} = \frac{\alpha_i}{n_t}, \\ \alpha_i = n_t \hat{r}_i, \quad \beta_i = n_t - \alpha_i, \end{aligned} \tag{8}$$

where $n_t = \alpha_i + \beta_i$ represents our belief on the confidence of the prior distribution estimation and a larger n_t will result in a smaller variance on the distribution estimation. Given n_t, both α_i and β_i can be derived from Eq. 8.

5.2 Online Monitoring and Estimation

As described in Algorithm 3, in the online phase, we update the distribution of the *trust score* with real-time measurements. We monitor the status of DERs (compromised or not) by comparing their readings with the estimated readings from a physics-guide GPR exploiting external data (e.g., weather), DER physical model, and real-time neighboring DER readings.

5.2.1 Physics-guided Gaussian process regression. We use the reported power generations from PV systems as the DER measurements and denote the generation of PV system i at time t as $\hat{P}_{D,i}(t)$. The true power generation of the PV system i is denoted as $m_i(t)$. We assume that $\hat{P}_{D,i}(t) \sim \mathcal{N}(m_i(t), \sigma_i^2(t))$ in the attack free scenario. Here $\sigma_i(t)$ is the standard deviation of $\hat{P}_{D,i}(t)$. Based on the distances between PV systems, each PV system i has a set of neighboring PV systems denoted as $E_i = \{j | d(i, j) < d_0; j \neq i; j \in N\}$.

$d(i,j)$ represents the distance between PV system i and j, and d_0 is the distance threshold. $E_{i,j}$ for $1 \le j \le k_i$ is the j-th element in E_i, i.e., the j-th neighbor of PV system i, and $k_i = |E_i|$ is the number of neighbors of PV system i. Since nearby PV systems share similar weather conditions, their performance will also follow similar patterns. Therefore, we can use the neighboring power generations to predict the power generation of PV system i.

Since the neighboring PV systems can be compromised, the prediction results may be biased with the tampered neighboring reports. Therefore, besides the neighboring data, we also combine the physical model predictions to enhance the prediction accuracy. Denote the power generation of PV system i derived from the physical model as $\tilde{P}_{D,i}(t)$. We assume that $\tilde{P}_{D,i}(t) \sim \mathcal{N}(m_i(t), \epsilon_i^2(t))$. $\epsilon_i(t)$ is the standard deviation of $\tilde{P}_{D,i}(t)$. Given the PV system manufacturing model, the solar irradiation $G_i(t)$ and the temperature $T_i(t)$ at PV system i's location at time t, the output voltage $V_i(t)$ and the output current $I_i(t)$ of PV system i at time t can be determined with Eq. 3 and Eq. 4, and thus the output power $\tilde{P}_{D,i}(t) = V_i(t)I_i(t)$ (Algorithm 3, line 3).

Based on the settings above, we have (Algorithm 3, line 4)

$$\mathbf{Z}_i(t) = (\hat{P}_{D,i}(t), \tilde{P}_{D,i}(t), \hat{P}_{D,E_{i,1}}(t), \tilde{P}_{D,E_{i,1}}(t), \ldots, \hat{P}_{D,E_{i,k_i}}(t), \tilde{P}_{D,E_{i,k_i}}(t))^T, \quad (9)$$

as the vector of the reported and the physics-based power generations for PV system i and its neighbors. For each epoch t over a day, $\mathbf{Z}_i(t)$ is collectively modeled as a Gaussian Process $\mathcal{N}(\mu_i(t), \Sigma_i(t))$. Here, $\mu_i(t)$ is the expectation of $\mathbf{Z}_i(t)$, i.e.,

$$\mu_i(t) = (m_i(t), m_i(t), m_{E_{i,1}}(t), m_{E_{i,1}}(t), \ldots, m_{E_{i,k_i}}(t), m_{E_{i,k_i}}(t))^T,$$

and $\Sigma_i(t)$ is the covariance matrix of $\mathbf{Z}_i(t)$. To parameterize the covariance matrix and incorporate performance similarities induced by physical distances, the covariance between any two elements of $Z_i(t)$ is defined based on $d(k,j)$ values by Gaussian kernel functions

$$Cov(\hat{P}_{D,k}(t), \hat{P}_{D,j}(t)) = \sigma(t)^2 \exp\left(-\frac{d(k,j)^2}{2h_1(t)^2}\right), \quad (10)$$

$$Cov(\tilde{P}_{D,k}(t), \tilde{P}_{D,j}(t)) = \epsilon(t)^2 \exp\left(-\frac{d(k,j)^2}{2h_2(t)^2}\right), \quad (11)$$

$$Cov(\hat{P}_{D,k}(t), \tilde{P}_{D,j}(t)) = \alpha(t)\sigma(t)\epsilon(t) \exp\left(-\frac{d(k,j)^2}{2h_3(t)^2}\right). \quad (12)$$

Here, $h_1(t)$, $h_2(t)$ and $h_3(t)$ are the length scales of the covariance functions adjusting the impact of PV system distances to their correlations. To simplify the mathematical model, here we assume all PV systems have the same standard deviations, that is, $\sigma_i(t) = \sigma(t)$ and $\epsilon_i(t) = \epsilon(t)$ for all $i \in \mathbf{N}$. Note that this assumption can be flexibly relaxed if the PV systems exhibit very large differences in variances. Since the physical model prediction from PV system k does not directly connect to the reported reading from PV system j, a parameter $\alpha(t)$ is used to scale the effect of their correlation.

During the training phase, we use historical data to estimate the hyper-parameters $\theta_i(t) = (\mu_i(t), \alpha(t), \sigma(t), \epsilon(t), h_1(t), h_2(t), h_3(t))$ by maximizing the log likelihood (Algorithm 3, line 5):

$$\mathcal{L}_i = -\frac{1}{2}\sum_{s=1}^{n_{sp}} \mathbf{Z}_{i,s}^T \Sigma_i \mathbf{Z}_{i,s} - \frac{1}{2} n_{sp} \times \log(|\Sigma_i|) - \frac{n_{sp}}{2} \times \log(2\pi), \quad (13)$$

where n_{sp} indicates the number of historical samples, and $\mathbf{Z}_{i,s}$ is the s-th sample..

During the prediction phase, the power generation of PV system i, $\bar{P}_{D,i}(t)$, is estimated with (Algorithm 3, line 13):

$$\bar{P}_{D,i}(t) = m_i(t) + K_i^{PZ}(t)(K_i^{ZZ}(t))^{-1}(\mathbf{Z}_{-i}(t) - \mu_{-i}(t)), \quad (14)$$

and the variance of the estimation, $\bar{K}_i(t)$, is

$$\bar{K}_i(t) = K_i^{PP}(t) - K_i^{PZ}(t)(K_i^{ZZ}(t))^{-1}(K_i^{PZ}(t))^T, \quad (15)$$

in which

$$\mathbf{Z}_{-i}(t) = (\tilde{P}_{D,i}(t), \hat{P}_{D,E_{i,1}}(t), \tilde{P}_{D,E_{i,1}}(t), \ldots, \hat{P}_{D,E_{i,k_i}}(t), \tilde{P}_{D,E_{i,k_i}}(t))^T, \quad (16)$$

$$\mu_{-i}(t) = (m_i(t), m_{E_{i,1}}(t), m_{E_{i,1}}(t), \ldots, m_{E_{i,k_i}}(t), m_{E_{i,k_i}}(t))^T, \quad (17)$$

$$K_i^{PZ}(t) = Cov(\hat{P}_{D,i}(t), \mathbf{Z}_{-i}(t)), \quad (18)$$

$$K_i^{ZZ}(t) = Cov(\mathbf{Z}_{-i}(t), \mathbf{Z}_{-i}(t)), \quad (19)$$

$$K_i^{PP}(t) = Cov(\hat{P}_{D,i}(t), \hat{P}_{D,i}(t)). \quad (20)$$

Algorithm 3 Online Trust Score Estimation

1: **function** GPRTraining(DER datasheets and locations, historical weather information $\{G_i\}$, $\{T_i\}$, historical DER measurements $\{\hat{P}_{D,i}\}$)
2: **for** $i \in \mathbf{N}$
3: Solve V_i and I_i with Eq. 3 and Eq. 4, and thus $\tilde{P}_{D,i} = V_i I_i$;
4: Construct Z_i according to Eq. 9;
5: $\theta_i = \arg\max_\theta \mathcal{L}_i$;
6: Calculate $\bar{K}_i(t)$, $K_i^{PZ}(t)$ and $K_i^{ZZ}(t)$ according to Eq. 15, Eq. 18 and Eq. 19, respectively;
7: **end for**
8: **return** $\{\theta_i(t)\}$, $\{\bar{K}_i(t)\}$, $\{K_i^{PZ}(t)\}$ and $\{K_i^{ZZ}(t)\}$;
9: **end function**
10:
11: **function** OnlinePhase($\{G_i(t)\}$, $\{T_i(t)\}$, $\{\hat{P}_{D,i}(t)\}$, $\{\theta_i(t)\}$, $\{\bar{K}_i(t)\}$, $\{K_i^{PZ}(t)\}$, $\{K_i^{ZZ}(t)\}$)
12: **for** $i \in \mathbf{N}$
13: Calculate the estimated DER measurement $\bar{P}_{D,i}(t)$ according to Eq. 14;
14: Calculate the statistic $t_i(t)$ according to Eq. 21;
15: Apply Wald test to $t_i(t)$ and obtain the trust score $s_i(t)$;
16: **end for**
17: **return** $\{s_i(t)\}$;
18: **end function**

5.2.2 Trust score in online phase. To decide whether the PV system i is under attack, the estimation $\bar{P}_i(t)$ is validated with the reported generation $\hat{P}_i(t)$ through Wald test (Algorithm 3, line 14-15). Since $\hat{P}_{D,i}(t)|Z_{-i}(t) \sim \mathcal{N}(\bar{P}_{D,i}(t), \bar{K}_i(t))$ in attack-free cases, the statistic

$$t_i(t) = (\hat{P}_{D,i}(t) - \bar{P}_{D,i}(t))/\sqrt{\bar{K}_i(t)} \quad (21)$$

follows the asymptotic z distribution. When the p-value corresponding to $t_i(t)$ is low, it is unlikely that the PV system is *not* under attack. We set up a threshold τ such that when the p-value is below τ, we believe the PV system is compromised and its online trust score $s_i = 0$. Otherwise, $s_i = 1$.

Since an adversary is prone to attack PV systems close to each other, the neighboring PV systems are also at risk when a PV system

is subject to attack. In our GPR, the estimations of the PV systems will be deviated if the neighboring PV system is compromised. Therefore, our *trust score* estimation framework can catch such potential vulnerabilities. On the other hand, thanks to the physical model, the impacts of compromised neighboring PV systems are alleviated to avoid the degradation of the grid performance.

5.3 Trust Score Update

Given an observation $s_i(t)$, the posterior distribution of r_i can be updated through the following exponentially weighted moving average of the prior and online trust score:

$$\begin{aligned} \alpha_i &\leftarrow (1-\lambda)\alpha_i + \lambda s_i(t), \\ \beta_i &\leftarrow (1-\lambda)\beta_i + \lambda(1-s_i(t)). \end{aligned} \tag{22}$$

Such an update can better utilize historical observations and facilitate the detection of less obvious attacks [18]. λ is a parameter to balance the prior knowledge and the new observations. A larger λ will make the *trust score* more sensitive to attacks while a smaller λ will result in a more stable *trust score*. Recall that $\hat{r}_i = E(r_i) = \alpha_i/n_t$ and $n_t = \alpha_i + \beta_i$, the posterior expectation of the *trust score* is then (Algorithm 1, line 6)

$$\hat{r}_i \leftarrow \frac{(1-\lambda)n_t\hat{r}_i + \lambda s_i(t)}{(1-\lambda)n_t + \lambda}. \tag{23}$$

6 RISK-AWARE DER MANAGEMENT

To maintain the normal operations of the grid, we integrate the estimated *trust scores* into the coordination of DERs in the transmission network by introducing a resilience term to balance the overall grid performance and the security risks derived from the *trust scores*. Depending on different grid operations, the optimization formulation may change. For illustration purpose, in this paper, we showcase how the *trust score* is utilized to maintain the voltage profile. Other tasks can be formulated in a similar manner.

To guarantee the grid stability or dispatch energies economically, the control center may want to set the grid voltages at a certain point. To force the voltage to the set point, the control center sends requests to PV systems to inject/absorb certain amounts of reactive powers to/from the grid. To decide the most preferrable action of each PV system, the control center calculates the optimal power flow (OPF). Nevertheless, when a PV system is compromised, the tampered power generation reports from it may mislead the decision of the control center. Moreover, the compromised PV systems may not follow the requests from the control center and thus, the grid cannot be correctly controlled. Therefore, to mitigate the impacts of tampered power generation reports, we use the estimated measurement from GPR $\bar{P}_{D,i}$ for OPF. Meanwhile, we introduce a resiliency term in which the assigned amount of the reactive power injection/absorption for each PV system $\mathbf{Q}_{D,i}$ is weighted with its *trust score* estimation $\hat{r}_i$. By doing so, we intend to assign heavier tasks to those trustworthy PV systems and balance the possible impacts caused by the malfunctioning of compromised PV systems. The objective function of OPF is formulated as follows:

$$\min_{\mathbf{Q}_D} ||\mathbf{V} - \mathbf{V}_{ref}||_2 - \eta \mathbf{r}^T \mathbf{Q}_D, \tag{24}$$

$\mathbf{V} = (V_1, V_2, \ldots, V_{n_b})$ is the vector of the voltage amplitudes of each bus. $\mathbf{V}_{ref}$ denotes the vector of desired values of the voltages. $\delta = (\delta_1, \delta_2, \ldots, \delta_{n_b})$ denotes the vector of the voltage angles of each bus. $\mathbf{r}$ is the vector of *trust scores* and $\mathbf{Q}_D$ is the vector of the amounts of reactive powers requested from PV systems. η is a penalty coefficient to balance the following two terms. The first term forces the grid voltage to the set voltage profile, and the second term enhances the grid resilience and limits the impacts of PV systems with low *trust scores*. By doing so, we increase the resilience of the grid by making a trade-off between grid performance and security.

There are several physical limits and constraints characterizing the system. All voltage magnitudes and angles are bounded by an upper and a lower limit to guarantee the stability of the grid:

$$\begin{aligned} 0.98 \le V_i \le 1.02, \\ -\frac{\pi}{2} \le \delta_i \le \frac{\pi}{2}. \end{aligned} \tag{25}$$

Since we consider a balanced system model, we have the following power balance constraints:

$$\begin{aligned} P_{G,i} + \bar{P}_{D,b_i} - P_{L,i} - V_i \sum_j V_j \left(G_{ij}\cos\delta_{ij} + B_{ij}\sin\delta_{ij}\right) = 0, \\ Q_{G,i} + Q_{D,b_i} - Q_{L,i} - V_i \sum_j V_j \left(G_{ij}\sin\delta_{ij} - B_{ij}\cos\delta_{ij}\right) = 0, \end{aligned} \tag{26}$$

$P_{G,i}, Q_{G,i}$ are the active and the reactive power injection from generator and $P_{L,i}, Q_{L,i}$ are the active and the reactive power consumed by the load at the bus i. All the 4 variables can be measured by smart meters. $\bar{P}_{D,b_i}$ and Q_{D,b_i} are the active and reactive powers of the PV system $b_i (b_i \in \mathbf{N})$ located at bus i. If no PV system is installed at the bus, $\bar{P}_{D,b_i}$ and Q_{D,b_i} will be 0. G_{ij} and B_{ij} are the real and the imaginary components of $\tilde{Y}_{ij}$. $\delta_{ij} = \delta_i - \delta_j$ is the voltage angle difference between bus i and j. Moreover, the apparent power of each PV system should not exceed its capacity:

$$P_{D,b_i}^2 + Q_{D,b_i}^2 \le S_{D,b_i}^2, \tag{27}$$

where S_{D,b_i} is the nominal power of PV system b_i.

We summarize the optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{Q}_D} \quad & ||\mathbf{V} - \mathbf{V}_{ref}||_2 - \eta \mathbf{r}^T \mathbf{Q}_D, \\ \text{s.t.,} \quad & Eq.\ 25, 26, 27 \end{aligned} \tag{28}$$

To solve the highly non-linear problem Eq. 28, we adopt a heuristic algorithm - Differential Evolution (DE) [6, 25]. DE improves the solution of the optimization problem iteratively by generating offspring candidates from mutations of the parent candidate solutions and selecting the better ones from the offspring and parent candidates. DE is easy to implement and converges fast because the mutation is generated based on the difference between candidates instead of random generation. Denote the number of populations as NP and the maximum number of generations as NG. The mutation factor F is the weight of the difference between two random candidates and functions like the learning rate. The crossover rate CR decides the probability that a crossover operation is performed. The implementation of DE is illustrated as follows:

(1) **Initialization:** Set the index of generation $g = 0$. A set of candidates $\{\mathbf{q}_k^0 | k \in \{1, \ldots, NP\}\}$ are initialized as random vectors between the lower and the upper bound of $\mathbf{Q}_D$, i.e., -S_D and S_D.

(2) **Mutation:** Increase the generation index $g \leftarrow g+1$. At the generation g, for each candidate, an offspring is generated from three candidates, $\mathbf{q}_{rk1}^{g-1}$, $\mathbf{q}_{rk2}^{g-1}$ and $\mathbf{q}_{rk3}^{g-1}$, randomly chosen from the candidate set: $\mathbf{u}_k = \mathbf{q}_{rk1}^{g-1} + F(\mathbf{q}_{rk2}^{g-1} - \mathbf{q}_{rk3}^{g-1})$.
(3) **Crossover:** For each element in $\mathbf{u}_k$, a random number $cr_{rand} \in (0,1)$ is generated. If $cr_{rand} > CR$, the element is replaced with the corresponding element in $\mathbf{q}_k^{g-1}$.
(4) **Selection:** Given $\mathbf{q}_k^{g-1}$ and $\mathbf{u}_k$, we derive $\mathbf{V}$ and δ with Matpower [29], and the results are evaluated w.r.t Eq. 28. The constraint violation is handled through the Superiority of feasible solutions (SF) method proposed in [10]. From $\mathbf{q}_k^{g-1}$ and $\mathbf{u}_k$, we choose the one with no or less violation that minimizes the objective function in Eq. 24 as the candidate for the next generation $\mathbf{q}_k^g$.
(5) The mutation, crossover and selection processes are iterated until the maximum number of generation is reached. From all candidates, we choose the one with no or the smallest violation that minimizes Eq. 24 as the solution of Eq. 28, $\mathbf{Q}_D$.

7 EXPERIMENTAL RESULTS

7.1 Dataset and Simulation Setup

We use the 39-bus system shown in Figure 2 as the test case. The PV systems are installed at the buses connected to loads but not connected to the generators, which results in 12 PV systems in total. The PV system indexes and their corresponding bus indexes are summarized in Table 1. The PV systems are divided into 3 groups with 4 PV systems in each group. Group 1 includes PV system 1, 4, 7, 10, group 2 includes PV system 2, 5, 8, 11, and group 3 includes PV system 3, 6, 9, 12. For PV systems in the same group, they are the neighbors of each other.

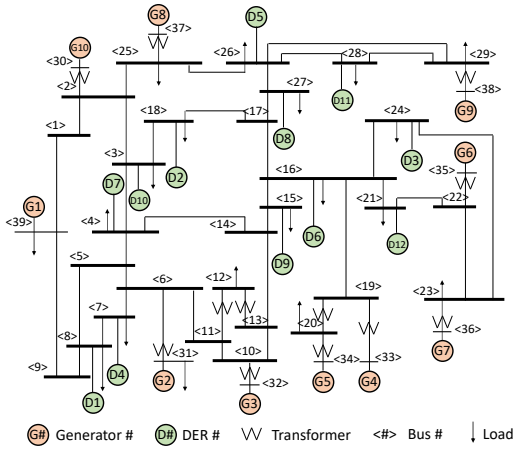

Figure 2: Single line diagram of the modified IEEE 39-bus with 12 PV installed.

PV	1	2	3	4	5	6	7	8	9	10	11	12
bus	8	18	24	7	26	16	4	27	15	3	28	21

Table 1: PV indexes and their corresponding bus indexes.

Due to the lack of real-time fine-grained load and PV generation data, we use synthetic data for simulation. Similar to [3], we use the load data of New York (NY) state in Nov, 2019 from the NYISO to derive the power demands of loads in the 39-bus system. There are 11 control areas in the NYISO map while there are 21 loads in the 39-bus system. Therefore, for each load in the 39-bus system, we randomly choose 2 control areas from the NYISO map and use the sum of their load data as the active power demand of the load in the 39-bus system. The synthetic data generation process is detailed in [3]. Since the historical reactive power data is not available, we generate the reactive powers for the loads by assuming a constant power factor $PF = 0.9$.

To generate the PV power generation data, we utilize the weather information from the National Solar Radiation Database (NSRDB) [22]. NSRDB offers synthetic data for half-hourly solar radiation measurements and meteorological data with a granularity of about 4km, i.e., 0.04°in latitude/longitude. We use global horizontal irradiance (GHI) and temperature measurements for the whole year of 2019. The locations of the 12 PV systems are selected from the California state, which is rich in the solar energies. The 3 groups of PV systems are located near (-81.58°, 30.5°), (-81.78°, 30.25°), and (-81.30°, 30.61°), respectively. We assume a penetration level of 15% and the nominal power of the PVs are set as 15% of the initial apparent powers of the loads in the 39-bus system. We use SunPower SPR-415E-WHT-D as the manufacturing model of PV systems. GHI and temperature measurements are input to Simulink [11] to generate the PV power generation data.

Based on the assumptions made in the threat model (Section 4.2), the attacks are performed on the 3 PV systems with the lowest offline trust scores. The offline trust scores of PV systems are presented in Figure 3, and the attack victims are PV system 1, 2 and 10. At each epoch, with a probability of 0.5, the attacks are performed by adding random numbers between 10 and 20 to the power generation data of the victim PV systems.

7.2 Simulation Results

7.2.1 Trust score in the offlline phase. In the simulation, the vulnerability score of PV system i, i.e., VS_i, is randomly generated. We calculate the offline trust scores of PV systems according to Eq. 7 with $n_t = 10$ and depict them in Figure 3. PV system 6 and 10, corresponding to bus 16 and 3, have the highest 2 NEC (0.26 and 0.34), which agrees with our intuitions that buses with higher degrees in graph theory are supposed to carry more power flows. On the other hand, PV system 10 also has a high VS value (0.54), and thus its offline trust score is the lowest among all the PV systems (0.82). This meets our assumption that a PV system located at a critical place with poor security implementations is more attractive to adversaries. Furthermore, although PV system 2 does not have a high NEC (0.17), it has the highest VS (0.72), i.e., the poorest security implementations. Therefore, PV system 2 is easy to be compromised and have a low overall offline trust score 0.88.

7.2.2 Physics-guided GPR evaluation. Since the power generations from PVs heavily rely on the solar irradiation, a GPR is trained for every half hour from 9 am to 16:30 pm when the solar irradiation is abundant, thus resulting in 16 GPR models for each PV system. In the training phase, 80% of the PV power generation data is used as the training data. To evaluate the performance of trustworthiness evaluation capability of the physics guided GPR, for each PV system, we use the mean of the relative errors $((Estimation - Truth)/Truth)$

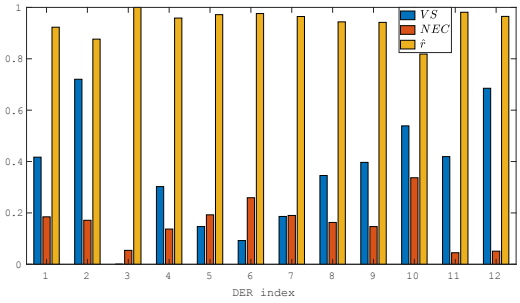

Figure 3: The offline trust score of each PV system.

at each time epoch as the performance metric, and the box plot of the average relative errors for each PV system is presented in Figure 4. When there's no attack, the average relative errors are within 5% except few outliers. Recall that the estimation accuracy depends on it's neighbor readings and physical model prediction. Therefore, when attack appears, the prediction errors increase for all the PV systems in group 1 (PV system 1, 4, 7, 10) as half of the PV systems (PV system 1 and 10) are compromised. On the other hand, for the PV systems in group 2 (PV system 2, 5, 8, 11), in which only PV system 2 is compromised, the performance of PV system 2 does not change because its neighbors are not attacked, and thus the predictions from GPR are the same. Besides, the performance of other PV systems in group 2 drops only slightly compared with the attack-free situation. This proves that the proposed GPR has the capability of recovering the true power generations of PV systems when a moderate number of PV systems are compromised.

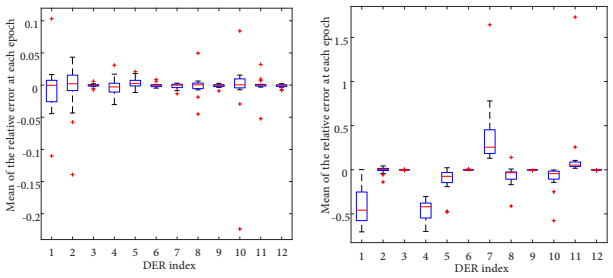

Figure 4: The box plot of the average relative errors at each epoch for PV systems. (a) depicts the results without attacks and (b) depicts the results with attacks.

When performing the Wald test, the threshold $\tau = 1 \times 10^{-5}$. An attack is alarmed if the online trust scores of any PV systems are equal to 0. The test achieves an accuracy ($correct/total$) of 98.75%, a false negative rate ($false\ negative/total\ positive$) of 0%, and a false positive rate ($false\ positive/total\ negative$) of 2.67%. Therefore, our online trust score estimation is capable of correctly evaluating the dynamic trustworthiness of the DERs.

7.2.3 The dynamics of trust scores. Here we evaluate the impact of attacks on the trust scores of the attacked PV systems and their neighbors as shown in Figure 5. The vertical red dashed line marks the epochs that a PV system is attacked over a day. Due to the superior performance of the online trust score estimation, we assign more weight to s_i and choose $\lambda = 0.5$ in Eq. 22. We select two PV systems from each group and present their trust scores $\hat{r}_i$ on day 1. According to the figure, we observe that the proposed trust score estimation framework can accurately capture the dynamic status of PV systems. The trust scores of the attacked PV systems (PV system 1 and 2) decrease at the epochs with attacks and increase at the epochs without attacks. Since two PV systems are attacked in group 1, the attacker is likely to attack the remaining ones. Thus the trust score of PV system 7 decreases as well. On the other hand, only PV system 2 is attacked in the group 2, thus PV system 5 is less affected and has a relatively high trust score. The trust scores of PV system 3 and 9 keep increasing because PV systems in the group 3 are not attacked.

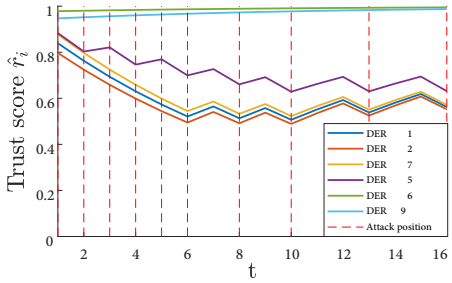

Figure 5: The dynamics of the trust scores on day 1.

7.2.4 Evaluation of RADM. With the trust scores, we optimize Eq. 28 with $\eta = 0.01$. Through several trials, we set $NP = 50, NG = 100, F = 0.7$ and $CR = 0.9$. The objective voltages of all buses are set at 1 p.u.. We use the Mean Square Errors (MSEs) between the reference voltage profile $\mathbf{V}_{ref}$ and the voltages $\mathbf{V}$ from the solution of Eq. 28 as the performance metric. In the basic case, the optimization is run without the second term in Eq. 28, i.e., with the objective to minimize $||\mathbf{V} - \mathbf{V}_{ref}||_2$ only. Both the basic case and RADM are examined with and without the presence of attacks and the results are presented in Figure 6. We use the MSE of the basic case without attacks as the baseline. As shown in Figure 6a, an average MSE of 5.42×10^{-4} is achieved in the baseline. Compared with the baseline, the attacks result in an average MSE of 6.39×10^{-4} in the basic case, which increases the average MSE by 17.97% (Figure 6c). On the other hand, in RADM, when no attack appears, the introduction of the resilience term, i.e., $-\eta \mathbf{r}^T \mathbf{Q}_D$ in Eq. 28, leads to an average MSE of 5.78×10^{-4}, which increases the average MSE by 6.56% (Figure 6b). When there are attacks, the average MSE is 5.77×10^{-4} (Figure 6d). Therefore, by introducing the resilience term, RADM resists to the attacks at the cost of slightly degrading the performance compared with the baseline. In the worst situation, a maximum MSE of 6.19×10^{-4} is observed in the baseline, which is increased by 61.44% with a maximum MSE of 1×10^{-3} when attacks occur. On the other hand, RADM achieves maximum MSEs of 6.36×10^{-4} without attacks and 6.46×10^{-4} with attacks, which are only 2.63% and 4.15% worse compared with the baseline. Thus, RADM could perform even better and significantly limit the impact of attacks in the worst situations.

8 CONCLUSIONS

In this paper, we develop a system-level DER management framework, RADM, which is capable of identifying DER risk levels and maintaining the grid performance even when DERs are subject to attacks. We propose to use *trust scores* to evaluate the trustworthiness of DERs and a trust score estimation method is developed. The

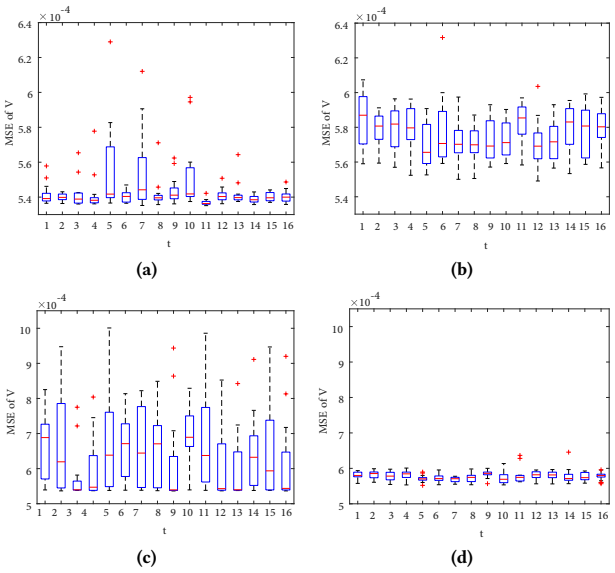

Figure 6: The box plot of the MSEs at each epoch. (a) the performance of the basic case without attacks; (b) the performance of RADM without attacks; (c) the performance of the basic case with attacks; and (d) the performance of RADM with attacks.

method estimate the *trust scores* by generating a general belief of the trust scores in the offline phase and then updating the belief with real-time data in the online phase. To mitigate the attack impacts on DERs, we formulate the grid decision making process as a optimization problem balancing the grid performance and the security risks derived from the *trustscores*. Through simulations, we use PV systems as a case study of DERs and demonstrate the capability of the *trust scores* of capturing the dynamic status of DERs as well as RADM's capability of mitigating the attack impacts with only slight degradation of the grid performance. In the future work, we will investigate developing a more robust GPR adapting to varying PV power generation patterns across a day to enhance the scalability of the proposed method.

ACKNOWLEDGMENTS

This work is partially supported by National Science Foundation (CNS-1818500) and CyberFlorida Collaborative Seed Award Program.

REFERENCES

[1] 2015. IEEE Standard Cybersecurity Requirements for Substation Automation, Protection, and Control Systems. *IEEE Std C37.240-2014* (2015), 1–38. https://doi.org/10.1109/IEEESTD.2015.7024885

[2] 2018. IEEE Standard for Interconnection and Interoperability of Distributed Energy Resources with Associated Electric Power Systems Interfaces. *IEEE Std 1547-2018 (Revision of IEEE Std 1547-2003)* (2018), 1–138. https://doi.org/10.1109/IEEESTD.2018.8332112

[3] Olugbenga Moses Anubi and Charalambos Konstantinou. 2019. Enhanced resilient state estimation using data-driven auxiliary models. *IEEE Transactions on Industrial Informatics* 16, 1 (2019), 639–647.

[4] Anomadarshi Barua and Mohammad Abdullah Al Faruque. 2020. Hall Spoofing: A Non-Invasive DoS Attack on Grid-Tied Solar Inverter. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 1273–1290.

[5] Rojan Bhattarai, Sheikh Jakir Hossain, Junjian Qi, Jianhui Wang, and Sukumar Kamalasadan. 2018. Sustained system oscillation by malicious cyber attacks on distributed energy resources. In *2018 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 1–5.

[6] Partha P Biswas, Ponnuthurai N Suganthan, Rammohan Mallipeddi, and Gehan AJ Amaratunga. 2018. Optimal power flow solutions using differential evolution algorithm integrated with effective constraint handling techniques. *Engineering Applications of Artificial Intelligence* 68 (2018), 81–100.

[7] Cedric Carter, Ifeoma Onunkwo, Patricia Cordeiro, and Jay Johnson. 2017. Cyber security assessment of distributed energy resources. In *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*. IEEE, 2135–2140.

[8] Critical Infrastructure Cybersecurity. 2014. Framework for Improving Critical Infrastructure Cybersecurity. *Framework* 1, 11 (2014).

[9] Ricardo Siqueira de Carvalho and Danish Saleem. 2019. Recommended functionalities for improving cybersecurity of distributed energy resources. In *2019 Resilience Week (RWS)*, Vol. 1. IEEE, 226–231.

[10] Kalyanmoy Deb. 2000. An efficient constraint handling method for genetic algorithms. *Computer methods in applied mechanics and engineering* 186, 2-4 (2000), 311–338.

[11] Simulink Documentation. 2020. Simulation and Model-Based Design. https://www.mathworks.com/products/simulink.html

[12] Sasan Gholami, Sajeeb Saha, and Mohammad Aldeen. 2017. A cyber attack resilient control for distributed energy resources. In *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*. IEEE, 1–6.

[13] William Hupp, Adarsh Hasandka, Ricardo Siqueira de Carvalho, and Danish Saleem. 2020. Module-OT: A Hardware Security Module for Operational Technology. In *2020 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 1–6.

[14] Samuel Koebrich, Thomas Bowen, and Austen Sharpe. 2018. 2018 Renewable Energy Data Book. *U.S. Department of Energy (DOE), Office of Energy Efficiency & Renewable Energy (EERE)* (2018).

[15] Christine Lai, Patricia Cordeiro, Adarsh Hasandka, Nicholas Jacobs, Shamina Hossain-McKenzie, Deepu Jose, Danish Saleem, and Maurice Martin. 2019. Cryptography considerations for distributed energy resource systems. In *2019 IEEE Power and Energy Conference at Illinois (PECI)*. IEEE, 1–7.

[16] Zhen Li, Deqing Zou, Shouhuai Xu, Hai Jin, Hanchao Qi, and Jie Hu. 2016. VulPecker: an automated vulnerability detection system based on code similarity analysis. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*. 201–213.

[17] Bin Liu, Zhen Li, Xi Chen, Yuehui Huang, and Xiangdong Liu. 2017. Recognition and vulnerability analysis of key nodes in power grid based on complex network centrality. *IEEE Transactions on Circuits and Systems II: Express Briefs* 65, 3 (2017), 346–350.

[18] James M Lucas and Michael S Saccucci. 1990. Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics* 32, 1 (1990), 1–12.

[19] Michael Mylrea and Sri Nikhil Gupta Gourisetti. 2017. Blockchain for smart grid resilience: Exchanging distributed energy at speed, scale and security. In *2017 Resilience Week (RWS)*. IEEE, 18–23.

[20] James Obert, Patricia Cordeiro, Jay Johnson, Gordon Lum, Tom Tansy, Max Pala, and Ronald Ih. 2019. Recommendations for trust and encryption in DER interoperability standards. In *Tech. Report, Sandia National Laboratories*. SAND2019–1490.

[21] D Jonathan Sebastian, Utkarsh Agrawal, Ali Tamimi, and Adam Hahn. 2019. DERTEE: Secure distributed energy resource operations through trusted execution environments. *IEEE Internet of Things Journal* 6, 4 (2019), 6476–6486.

[22] Manajit Sengupta, Yu Xie, Anthony Lopez, Aron Habte, Galen Maclaurin, and James Shelby. 2018. The national solar radiation data base (NSRDB). *Renewable and Sustainable Energy Reviews* 89 (2018), 51–60.

[23] Saleh Soltan, Prateek Mittal, and H Vincent Poor. 2018. BlackIoT: IoT botnet of high wattage devices can disrupt the power grid. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 15–32.

[24] Pirathayini Srikantha and Deepa Kundur. 2015. A DER attack-mitigation differential game for smart grid security analysis. *IEEE Transactions on Smart Grid* 7, 3 (2015), 1476–1485.

[25] Rainer Storn and Kenneth Price. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* 11, 4 (1997), 341–359.

[26] Hongmei Tian, Fernando Mancilla-David, Kevin Ellis, Eduard Muljadi, and Peter Jenkins. 2012. *Detailed performance model for photovoltaic systems*. Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

[27] Durgadevi Velusamy and Ganesh Kumar Pugalendhi. 2019. Fuzzy integrated Bayesian Dempster–Shafer theory to defend cross-layer heterogeneity attacks in communication network of Smart Grid. *Information Sciences* 479 (2019), 542–566.

[28] Zhifang Wang, Anna Scaglione, and Robert J Thomas. 2010. Electrical centrality measures for electric power grid vulnerability analysis. In *49th IEEE conference on decision and control (CDC)*. IEEE, 5792–5797.

[29] RD Zimmerman and CE Murillo-Sánchez. 2020. Matpower [Software].

[30] Ioannis Zografopoulos and Charalambos Konstantinou. 2020. DERauth: A Battery-based Authentication Scheme for Distributed Energy Resources. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 560–567.