Synthetic PMU Data Creation Based on Generative Adversarial Network Under Time-varying Load Conditions

Xiangtian Zheng, *Student Member*, *IEEE*, Andrea Pinceti, *Member*, *IEEE*, Lalitha Sankar, *Senior Member*, *IEEE*, and Le Xie, *Fellow*, *IEEE*

***Abstract*—In this study, a machine learning based method is proposed for creating synthetic eventful phasor measurement unit (PMU) data under time-varying load conditions. The proposed method leverages generative adversarial networks to create quasi-steady states for the power system under slowly-varying load conditions and incorporates a framework of neural ordinary differential equations (ODEs) to capture the transient behaviors of the system during voltage oscillation events. A numerical example of a large power grid suggests that this method can create realistic synthetic eventful PMU voltage measurements based on the associated real PMU data without any knowledge of the underlying nonlinear dynamic equations. The results demonstrate that the synthetic voltage measurements have the key characteristics of real system behavior on distinct time scales.**

***Index Terms*—Synthetic phasor measurement unit Data, Generative adversarial networks, Neural ordinary differential equations.**

I. Introduction

OVER the past decade, thousands of phasor measurement units (PMUs) have been deployed in backbone transmission systems in North America and abroad. This enables improved monitoring and control of the power system dynamics at considerably higher resolutions than previously possible. Transient dynamic data recorded by PMUs are of particular value to the research community for distinct research interests such as real-time monitoring, control, and protection. Although machine learning (ML) based methods have been proposed for a wide range of tasks such as those in [1]-[4], the practical development of ML-based methods for real cases using real eventful PMU data is obstructed by limited data availability, which is mainly attributed to two reasons: ① the real operational data of power grids are typically confidential and mostly prevented from being publicly available because of strict policies regarding critical energy/electric infrastructure information; ② given the reliability and stability of power grids, high-impact events such as system-wide voltage oscillation are rarely observed in real PMU data, and even if such events are observed, they are not often labeled.

Therefore, it is critical for public researchers to create a massive amount of realistic eventful PMU data to train, test, and calibrate data-driven methods that can be applied to real cases. Although researchers have recently contributed to the creation of datasets based on large-scale realistic synthetic grid models [5] for analysis, such as macroscopic energy portfolio transitions [6], [7] and major event reproduction [8], the value of real eventful PMU data cannot be exploited by existing methods that generate data by simulation. Other recent studies have contributed to the development of ML-based methods for generating power system data, such as load profile generation [9], [10], renewable scenario generation [11], and eventful PMU generation [12], [13], and have proposed potential uses for synthetic PMU data, such as disturbance classification with improved accuracy [13], load forecasting, and optimal power flow [10]. However, several gaps remain in existing work regarding the creation of a massive amount of realistic large-scale eventful PMU data at multiple time scales and with arbitrary lengths. First, the prior success of PMU data generation methods in small-scale Institute of Electrical and Electronics Engineers (IEEE) standard systems may not meet the demand for synthetic data based on real PMU datasets. Second, the short horizon of synthetic data limits the generalization of their applications. Finally, the lack of incorporation of time-varying load conditions may undermine the fidelity of long-length synthetic PMU data. Moreover, researchers [14] recently demonstrated that the general state-of-the-art methods for generating time series [15]-[18] developed in the ML community are not capable of creating synthetic PMU time series with good diversity and fidelity. This is because of the high dimensionality of the data and the need to model physical-based constraints.

Manuscript received: December 3, 2021; revised: March 31, 2022; accepted: June 10, 2022. Date of CrossCheck: June 10, 2022. Date of online publication: XX XX, XXXX.

The work was supported by the National Science Foundation (No. OAC-1934675, No. ECCS-2035688, No. ECCS-1611301).

X. Zheng and L. Xie (corresponding author) are with the Department of Electrical and Computer Engineering at Texas A&M University, College Station, TX 77840, USA (e-mail: zxt0515@tamu.edu; le.xie@tamu.edu).

A. Pinceti was with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA, and he is now with Dominion Energy, Richmond, VA 23219, USA (e-mail: Andrea.Pinceti@asu.edu).

L. Sankar is with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA (e-mail: lsankar@asu.edu).

DOI: 10.35833/MPCE.2021.000783

To address these challenges, we propose a method for generating eventful PMU data based on limited real data that leverages generative adversarial networks (GANs) to create quasi-steady states for the power system under time-varying load conditions and utilizes neural ordinary differential equations (ODEs) to capture the transient behaviors of the system during voltage oscillation events. This method is potentially generalizable to other real power systems. We separately validate the fidelity of the synthetic load and voltage oscillation data from various perspectives.

The contributions of this paper are summarized as follows.

1) Generation of data-driven eventful PMU measurements. The proposed method for generating eventful PMU voltage measurements can create realistic-looking PMU streams that capture the patterns of load changes and system oscillations over distinct time scales, of which the fidelity and scalability are demonstrated for a large-scale real dataset.

2) Efficient data generation algorithm. The proposed method achieves an efficient learning process by decoupling distinct time scales separately and leveraging the low-rank property of high-dimensional datasets.

The remainder of this paper is organized as follows. Section II introduces the problem formulation for the task of creating synthetic PMU data using ML. Section III briefly reviews the basic concepts of the GAN and neural ODE model adopted in this study. Section IV proposes a method for creating eventful PMU data under time-varying load conditions. Section V presents a case study using a real dataset. Finally, Section VI draws conclusions and plans for future work.

II. Problem Formulation

In this section, we present mathematical formulations for the task of generating eventful PMU data. Here, we only have access to the power flow model of a large-scale real system and no knowledge of the dynamic model. We assume that the created multi-time-scale PMU measurements are a linear combination of the steady-state voltage and voltage oscillation, which are determined by the pattern of changes in the load and the nature of the system dynamics, respectively. Therefore, the task is separated into two subtasks: ① the generation of steady-state voltage measurements and ② the generation of voltage oscillation measurements. We further discuss the challenges and propose the corresponding instructions for the method design.

A. Generation of Steady-state Voltage Measurements

Consider a set of historical PMU measurements including voltage and current measurements. We denote the voltage measurement matrix $\boldsymbol{V}^{\mathrm{SS}}$ as:

$$\boldsymbol{V}^{\mathrm{SS}}=\begin{bmatrix} V_{1,1}^{\mathrm{SS}} & V_{1,2}^{\mathrm{SS}} & \cdots & V_{1,N}^{\mathrm{SS}} \\ V_{2,1}^{\mathrm{SS}} & V_{2,2}^{\mathrm{SS}} & \cdots & V_{2,N}^{\mathrm{SS}} \\ \vdots & \vdots & & \vdots \\ V_{M,1}^{\mathrm{SS}} & V_{M,2}^{\mathrm{SS}} & \cdots & V_{M,N}^{\mathrm{SS}} \end{bmatrix} \tag{1}$$

where $V_{i,j}^{\mathrm{SS}}$ is the voltage at PMU j at time $i\Delta T$, and ΔT is the sampling period; N is the number of PMUs; and M is the number of time steps.

We assume that the voltage measurements are collected when the system is in a quasi-steady state. The task for generating steady-state voltage measurements aims to develop a data creation algorithm using the real samples $\boldsymbol{V}^{\mathrm{SS}}$ such that the synthetic multichannel time-series data $\hat{\boldsymbol{V}}_{M'\times N}^{\mathrm{SS}}$, containing N measurement channels over M' arbitrary time steps, exhibit similar properties as those of the historical data, such as the slowly-varying pattern attributed to changes in the load.

B. Generation of Voltage Oscillation Measurements

We denote the voltage oscillation measurement matrix as $\boldsymbol{V}^{\mathrm{OS}}$ with the same definition, which is collected under eventful system conditions. We assume that $\boldsymbol{V}^{\mathrm{OS}}$ can be expressed by a linear combination of the equilibrium voltage $\overline{\boldsymbol{V}}^{\mathrm{OS}}$ and voltage oscillation $\tilde{\boldsymbol{V}}^{\mathrm{OS}}$.

$$\boldsymbol{V}^{\mathrm{OS}}=\tilde{\boldsymbol{V}}^{\mathrm{OS}}+\overline{\boldsymbol{V}}^{\mathrm{OS}} \tag{2}$$

The task for generating voltage oscillation measurements aims to learn the pattern of the voltage oscillation $\tilde{\boldsymbol{V}}^{\mathrm{OS}}$ using real samples $\boldsymbol{V}^{\mathrm{OS}}$ such that the created synthetic time-series data $\hat{\boldsymbol{V}}_{M'\times N}^{\mathrm{OS}}$, containing N measurement channels over M' arbitrary time steps, exhibit realistic properties such as the decaying periodic oscillation determined by the dynamic characteristics of the system and the low rank due to the high coherency throughout the system.

C. Challenges

Although we separate the task for generating PMU measurement data into two subtasks, two key challenges still need to be resolved for ML-based synthetic PMU data generation approaches: ① enabling an ML-based data generation method to efficiently learn from a high-dimensional dataset and ② guaranteeing that the created PMU data are meaningful in terms of complying with physical laws. The remainder of this subsection discusses our method for addressing these challenges and describes the resulting algorithm design.

1) Efficient Creation of High-dimensional Data

The dimensions of the time-series data M and N are nontrivial in the context of PMU data generation. A high dimensionality may render the training process intractable and degrade the performance of the generative algorithms. Therefore, the proposed method should address these challenges from both temporal and spatial perspectives. First, the proposed method can decompose a long time series into multiple time resolutions and separately learn the temporal correlations of distinct time scales. Second, the proposed method can reduce the order of high-dimensional measurements by utilizing existing low-rank characteristics, which are attributed to a strong spatial correlation.

2) Data Fidelity

As real PMU measurements comply with physical laws, data fidelity, one of the main criteria for synthetic data quality, is another challenge. It requires Kirchhoff's laws to be satisfied by the synthetic data at each snapshot and that the evolving synthetic time series follow the characteristics of the dynamics of the power system. For the first requirement, the proposed method can create synthetic load profiles and calculate synthetic voltage measurements via power flow simulation to automatically guarantee Kirchhoff's laws. For

the second requirement, the method can learn fast oscillation patterns using an ML model that embeds the ODE format.

III. Basic Concepts of GAN and Neural ODE Model

A. Review of GAN

GANs, first proposed in [19], have now arguably become one of the most popular and successful deep generative models in multiple fields and disciplines [20]-[22].

The two key models of a GAN, the generative model (generator) G and discriminate model (discriminator) D, are implemented by neural networks, which are iteratively updated by optimizing the objective function $\mathcal{J}$ as:

$$\min_G \max_D \mathcal{J} = \mathop{\mathbb{E}}_{x} (\ln(D(x))) + \mathop{\mathbb{E}}_{z} (\ln(1 - D(G(z)))) \quad (3)$$

where x and z are the real data samples and random noise sampled from a predefined distribution, respectively; and $\mathbb{E}(\cdot)$ is an expectation function.

Additionally, another variant of a GAN [23] implements conditional data generation by modifying the objective function to.

$$\min_G \max_D \mathcal{J} = \mathop{\mathbb{E}}_{x,y} (\ln(D(x,y))) + \mathop{\mathbb{E}}_{z,y} (\ln(1 - D(G(z,y)))) \quad (4)$$

where y is a label representing the category of interests.

B. Review of Neural ODEs

The neural ODE model [24] is widely used for time-series modeling and regression for irregular time series. It comprises two key components: a neural network and an ODE solver. Instead of specifying a discrete sequence of hidden layers, this model parameterizes the derivative of a state using a neural network f_{θ_f}. It can be trained by supervised learning to minimize a scalar-valued loss function $\mathcal{L}(s)$ as follows:

$$\min_{\theta_f} \mathcal{L}(s) = \frac{1}{t_1 - t_0} \sum_{t=t_0}^{t_1} \left\| \int_{t_0}^{t} f_{\theta_f}(s(\tau))\mathrm{d}\tau + s(t_0) - s(t) \right\|_2 \quad (5)$$

where $\int_{t_0}^{t} f_{\theta_f}(s(\tau))\mathrm{d}\tau + s(t_0)$ is the estimated state at time t; $s(t)$ is the result of measurements at time t; and f_{θ_f} is the function representing a neural network parameterized by θ_f, which indicates how the measurements evolve along the timeline.

IV. Proposed Method for Creating Eventful PMU Data Under Time-varying Load Conditions

We assume that the multi-time-scale eventful PMU measurements are a linear combination of steady-state voltage measurements and voltage oscillation measurements, which are determined by the pattern of changes in the slowly-varying load and the nature of the fast-varying system dynamics, respectively. With this assumption, we separate the eventful PMU data generation task into two subtasks. The first aims to create realistic time-varying load profiles and then estimate the steady-state voltage measurements via a power flow simulation based on the obtained system model. The second subtask aims to synthesize realistic voltage oscillation profiles that follow the periodic patterns of the real transient dynamics of the system. With such an instructive principle, a novel algorithm that generates two-stage PMU data using a GAN [19] and neural ODEs [24] is proposed, as shown in Fig. 1.

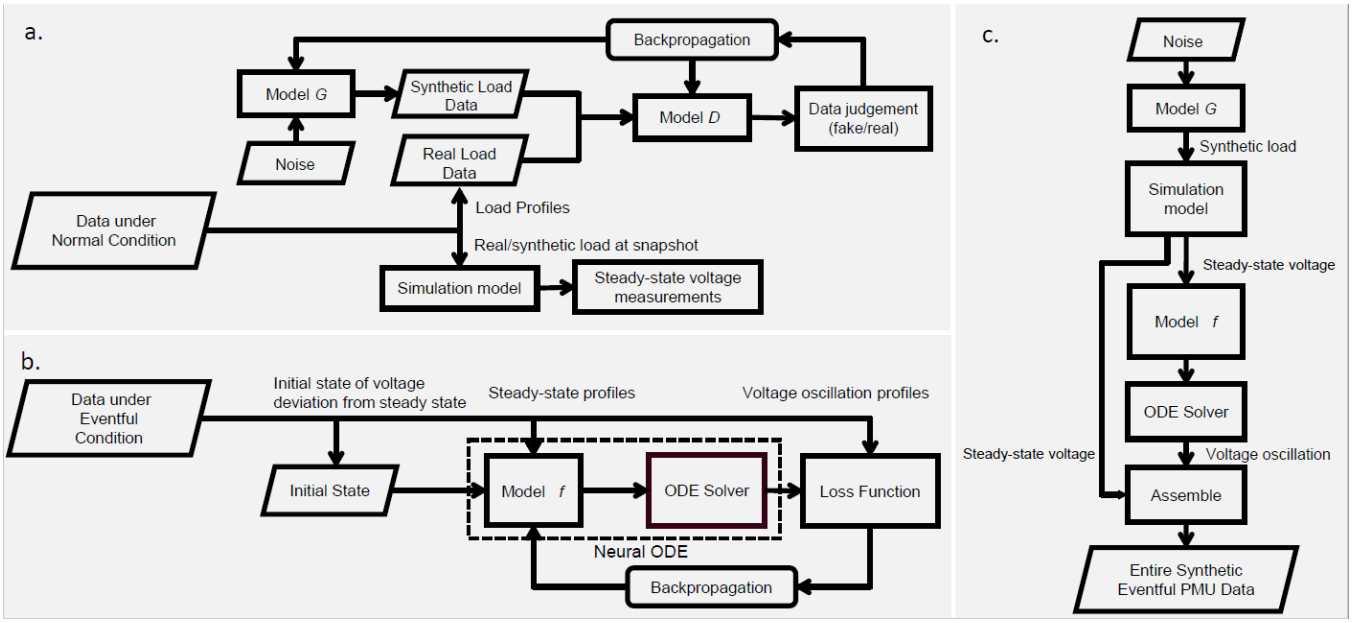

Fig. 1. Proposed method incorporating GAN and neural ODE models. (a) Training process of the GAN model for synthetic load data that is the input of simulation for steady-state voltage measurements. (b) Training process of the Neural ODE model for synthetic voltage oscillation measurements. (c) Generation of entire synthetic voltage measurements by the trained GAN and Neural ODE models.

In Fig. 1(a), the synthetic steady-state voltage measurements are simulated using synthetic load data generated by the trained multiresolution GAN model. In Fig. 1(b), the synthetic voltage oscillation measurements are generated by the neural ODE model that learns the system dynamics in the ODE format. Here, $f$is a neural network function. In Fig. 1 (c), the trained models G and f can be assembled to generate synthetic eventful PMU data. In the training process, the

GAN model is trained to create synthetic time-varying load profiles to estimate the simulation-based steady-state voltage measurements, whereas the neural ODE model is trained to generate a voltage oscillation with the limited real eventful voltage measurements as the training data. In the data creation process, we combine the well-trained models G and f to generate eventful PMU data with an arbitrary length under the given synthetic time-varying load conditions. The proposed method is feasible as long as the number of synthetic variables is less than the number of independent variables in the algebraic equations that are mainly derived from Kirchhoff's laws. In other words, this generation framework is compatible with the synthesis of voltage or current measurements. In this paper, we only show the case of voltage measurement generation to avoid verbosity.

The remainder of this section introduces the detailed algorithms for ① the generation of steady-state voltage measurements that consists of GAN-based load profile generation and simulation-based estimation of the steady-state voltage measurements and ② the generation of voltage oscillation measurements that leverages neural-ODE-based time-series learning.

A. Generation of Steady-state Voltage Measurements

The task for generating steady-state voltage measurements consists of two steps: ① the generation of a GAN-based multiresolution load profile [9]; and ② simulation-based estimation of the steady-state voltage under synthetic time-varying load conditions.

1) GAN-based Load Profile Generation

We use the algorithm for generating a multiresolution bus-level load profile proposed in [9]. This algorithm aims to develop a scheme to generate realistic time-series load data varying in length from a few minutes to a year and at varying resolutions from one sample per week to one sample per minute. We train independent generative models to capture the characteristics of these load profiles via a data down-sampling and aggregation process at different levels, which is summarized in the following steps.

1) Compute the power consumption of different load buses using PMU voltage and current measurements.

2) Down-sample the load data into multiple time scales and resolutions, including hour-long profiles at two samples per minute, week-long profiles at one sample per hour, and year-long profiles at one sample per week.

3) Train a generative model for the load profiles at each time scale and resolution, which is implemented by the conditional GAN in Algorithm 1, where ∇_{θ_D} and ∇_{θ_G} calculate the gradients with respect to parameters θ_D and θ_G, respectively, and *RMSProp* represents a root mean squared propagation function.

Algorithm 1: algorithm for generating GAN-based bus-level load profile

Require: historical load data at a certain time scale X, associated labels Y, random noise data Z, learning rate α, batch size m, initial parameter θ_D for the model D, and initial parameter θ_G for the model G

while θ_D and θ_G not converged

Sample batch $\{(x_i, y_i)\}_{i=1}^{m}$ from X and Y

Sample batch $\{(z_i, y_i)\}_{i=1}^{m}$ from Z and Y

#Update the model D using gradient descent

$$g_{\theta_D} \leftarrow \nabla_{\theta_D} \frac{1}{m}\left(-\sum_{i=1}^{m} D(x_i, y_i) + \sum_{i=1}^{m} D(G(z_i, y_i))\right)$$

$$\theta_D \leftarrow \theta_D - \alpha \cdot RMSProp(\theta_D, g_{\theta_D})$$

#Update the model G using gradient descent

$$g_{\theta_G} \leftarrow \nabla_{\theta_G} - \frac{1}{m}\sum_{i=1}^{m} D(G(z_i, y_i))$$

$$\theta_G \leftarrow \theta_G - \alpha \cdot RMSProp(\theta_G, g_{\theta_G})$$

end while

2) Simulation-based Estimation of Steady-state Voltage Measurements

Using the power flow simulation model accompanied by the dataset, we estimate the steady-state voltage measurements under certain load conditions by performing a power flow simulation at every time step. Given one synthetic load profile, the power flow simulation is repeatedly performed at each time step such that all system loads and the generation are scaled by the per-unit value of the load profile at the snapshot. Here, we admit that generation dispatch under different load conditions is simple without incorporating factors such as power markets and planned outages, which require further investigation but are outside the scope of this paper.

In summary, we generate steady-state voltage measurements in two steps. By leveraging a model that generates well-trained load profiles, we first generate a massive number of realistic load profiles during a certain time period that have a similar pattern but exhibit diversity. By assigning synthetic load profiles to the load buses in the simulation model and proportionally scaling the generation dispatch, we obtain a massive number of steady-state voltage measurements at different time scales and resolutions via power flow simulation.

B. Generation of Voltage Oscillation Measurements

Inspired by the data-driven system identification method SINDy [25], the method for learning the nonlinear dynamics consists of modular steps including decomposition, feature extraction, and time-series learning and leverages neural networks to learn the oscillation pattern of the extracted low-dimensional feature time series, as shown in Fig. 2.

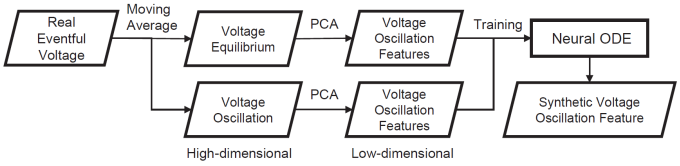

Fig. 2. Diagram of training the Neural ODE model for generating voltage oscillation measurements.

The details are summarized in the following steps and formally presented in Algorithm 2, where ∇_{θ_f} calculates the gradient with respect to parameters θ_f, f_{MA} is a moving average function that returns the average value and residual of time series in a moving window, and F_{OS} returns the intergal of a function over using an ODE solver.

Algorithm 2: algorithm for generating voltage oscillation measurements

Require: eventful voltage measurements V, reduced principle component analysis (PCA) approximation rank r, batch size m, learning rate α, loss function $\mathcal{L}$, and initial parameter θ_f for model f

Decomposition

$(\tilde{V}, \overline{V}) \leftarrow f_{\mathrm{MA}}(V)$

Dimension reduction

$\tilde{Z} \leftarrow PCA(\tilde{V}, r)$

$\overline{Z} \leftarrow PCA(\overline{V}, r)$

Train time-series learning model

while θ_f not converged

Sample batch $\{(\tilde{Z}_i, \overline{Z}_i)\}_{i=1}^{m}$ that are segments from $\tilde{Z}$ and $\overline{Z}$

Generate synthetic data

$\tilde{Z}_i^* \leftarrow F_{OS}(f_{\theta_f}, \tilde{Z}_i, \overline{Z}_i), i = 1, 2, ..., m$

$g_{\theta_f} \leftarrow \nabla_{\theta_f} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\tilde{Z}_i, \tilde{Z}_i^*)$

$\theta_f \leftarrow \theta_f - \alpha \cdot RMSProp(\theta_f, g_{\theta_f})$

end while

1) Decomposition

To decompose the original voltage measurements into the equilibrium voltage $\overline{V}$ and voltage oscillation $\tilde{V}$, the moving average method is first used to process the original voltage measurements, where the average voltage calculated in the moving window is defined as the equilibrium voltage and the residual is defined as the voltage oscillation.

2) Feature Extraction

To implement feature extraction, principal component analysis (PCA) method is used to process the voltage oscillation $\tilde{V}$ and equilibrium voltage $\overline{V}$ to obtain the reduced-order features $\tilde{Z}$ and $\overline{Z}$. Here, the underlying assumption is that the characteristics of $\tilde{Z}$ and $\overline{Z}$ have a one-to-one correspondence with the original measurements $\tilde{V}$ and $\overline{V}$. The PCA method uses the parameter r to determine the number of principal components to be retained, which also indicates the reduced rank of the approximated data after reconstruction. We select the r principal components with the highest variances as the feature time series such that these components can explain at least 95% of the variability in the original measurements.

3) Oscillation Time Series Modeling

We assume that the equilibrium voltage is uniquely determined by the load conditions. The task of generating voltage oscillations under time-varying load conditions is thus equivalent to generating voltage oscillations when the equilibrium voltage varies. Therefore, we train a neural ODE model f to learn the oscillation pattern of the low-dimensional features $\tilde{Z}$ at the corresponding equilibrium $\overline{Z}$.

In summary, given the voltage oscillation measurements calculated by the moving average method, we first perform order reduction to improve the computational efficiency and reduce the model complexity and then leverage the neural ODE model to learn the underlying dynamic behavior of the extracted feature time series. As the synthetic steady-state voltage measurements are within the varying equilibrium, we can create a massive number of voltage oscillations using the well-trained model f, of which the data creation process also requires the PCA mapping matrix for transformation.

V. Case Study

In this section, we demonstrate the proposed method using a large-scale real PMU dataset. We first show that the generated load profiles and steady-state voltage measurements are visually indistinguishable from the real samples and exhibit the same statistical properties. We also show the fidelity of the generated voltage oscillation measurements using a modal analysis.

A. Data Description and System Model

In this study, we use a large-scale real PMU dataset obtained from a major United States electricity utility company. This dataset was collected at a rate of 30 samples per second for three consecutive years from approximately 400 PMUs throughout the utility's territory and mainly contains voltage and current measurements. Furthermore, we have access to a large-scale power simulation model of the relevant network that contains more than 30000 buses and covers the utility's territory. The dataset provides the unique identifiers of the PMU buses that are consistent with the simulation model, thereby enabling the localization of the PMUs in the simulation model.

On the basis of the system topology and placement of the PMUs, we identify 12 fully monitored load buses, of which the load demand can be directly calculated by the positive-sequence complex current and voltage measurements. The load profiles reflect the periodic patterns of load changes at different time scales. The dataset also contains seven system-wide voltage oscillation events in the records, where only one weakly damped event lasted for approximately 2 hours and the others quickly vanished. The weakly damped event shows the shifting dominant modes of the system oscillation.

In the remainder of this section, we demonstrate the proposed method by generating voltage equilibrium profiles based on real load profiles and creating voltage oscillation profiles based on quickly and weakly damped events.

B. Data Processing and Model Training

The details of the data processing and model training for the two subtasks are introduced below. The configuration of the neural network model and the computational environment are presented in Appendix A.

1) Generation of Steady-state Voltage Measurements

Following Algorithm 1, we train the GAN model using the real load profiles of the 12 identified load buses, for which we set the batch size m as 32, the learning rate α as 10^{-4}, and the maximum number of training epochs to 50000. The configurations of models G and D are shown in Appendix A Table AI. In sequence, we create 1000 1-hour-long minute-level (per-unit) load profiles that represent various load changes over different time periods such as day or night, weekdays or weekends, and seasons. Given one per-unit load profile created as an input, we first scale all loads and generation in the simulation model to guarantee balanced supply and demand and then solve for the power flow

every time step to obtain the voltage measurements. This process is implemented using Python codes that use the ESA package [26] to interact with the PowerWorld simulator using its SimuAuto function.

2) Generation of Voltage Oscillation Measurements

To separate the equilibrium voltage and voltage oscillation profiles, the moving average method is used to process the voltage measurements of each voltage oscillation event in the dataset, where the size of the moving window is set to be 10 s. The order of the processed high-dimensional voltage measurements is reduced to 4 by PCA, as these 4 dominant components can explain more than 95% of the variability. We train model f on low-rank features, as instructed in Algorithm 2 (configuration of model f is shown in Appendix A Table AI), where we set the batch size m as 32, the learning rate α as 10^{-3}, and the maximum number of training epochs as 50000.

C. Synthetic Steady-state Voltage Measurements

The GAN model for generating synthetic load profiles is trained with the power measurements at the fully monitored load buses in the real dataset as the training data, with the aim of having a realistic and diverse pattern. The fidelity is validated by comparing its statistical characteristics with those of real profiles.

The generative models for the time-series load data are validated with statistical comparisons. The following two metrics are used to verify that the synthetic data capture the characteristics of the real data.

1) Wasserstein distance. The goal of model G is to learn a function that maps the known noise distribution to the distribution of real data. Training is successful when the distribution of the generated data matches that of real data. The Wasserstein distance is a measure of the distance between two distributions, and it can be used to quantitatively assess the closeness of the distributions of the generated and real data.

2) Power spectral density (PSD). An important characteristic of time-series load data is periodicity. Because loads are tied to the routines and behaviors of people, they have different recurring patterns. One approach to identify these periodicities is to examine the PSD of time-series data. Figure 3 shows the comparison of the PSDs of real and synthetic load profiles, where three peaks of PSD correspond to three typical periods of loads, namely, 12 hours, 24 hours, and 1 week. As observed, the two profiles match very closely, confirming that the generated data capture the periodic behavior of real data.

In sequence, we create 1000 1-hour-long minute-level (per-unit) load profiles that represent diverse load changes over different time periods such as daytime or nighttime, weekday or weekend, and seasons. Given one per-unit load profile as an input, we scale all loads and generation in the simulation model and solve for the power flow at every time step. Finally, we obtain the steady-state voltage measurements of 1000 different load conditions by repeating the simulation. To validate the synthetic voltage measurements, we compare the distributions of the real and synthetic 1-hour-long steady-state voltage angle measurements under different load conditions for a PMU, as shown in Fig. 4. This demonstrates that the synthetic voltage measurements are in good agreement with the real measurements, which is attributed to the fidelity and diversity of the synthetic load profiles. Note that the differences between the real and synthetic distributions might be caused by different settings for the magnitude of the voltage and the overly simple generation dispatch, which we will address in future work.

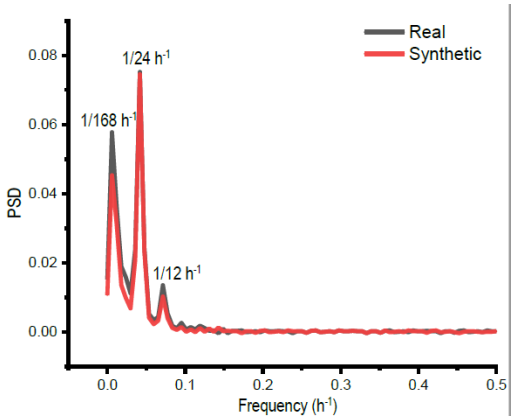

Fig. 3. Comparison of PSDs of real and synthetic load profiles.

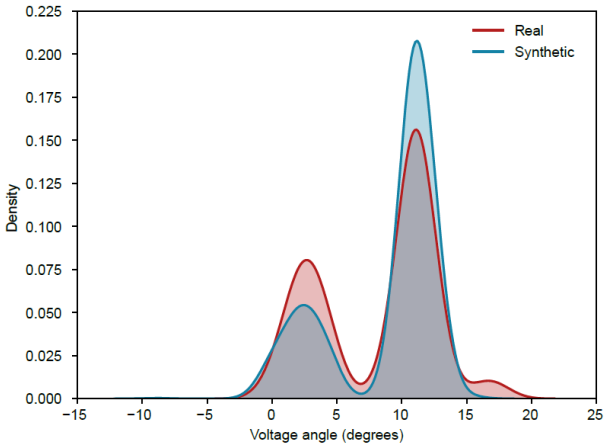

Fig. 4. Comparison of distributions of real and synthetic 1-hour-long steady-state voltage angle measurements under different load conditions for a PMU.

D. Synthetic Voltage Oscillation Measurements

The neural ODE model f for the voltage oscillation measurements is trained according to the details introduced in Section V-B. To demonstrate the learning capacity, the results for synthetic voltage oscillation data for two events of distinct duration are presented: a 10-second quickly damped oscillation event and a 2-hour weakly damped oscillation event.

We first train the neural ODE model with the voltage measurements in a 10-second-long event as the training dataset. The visual comparison in Fig. 5 demonstrates the fidelity and the flexibility of the length of the synthetic time series. More specifically, the first 10 s data of the synthetic time series (blue solid) validate the fidelity, whereas the following 5

s data show the flexibility of the length of the time series generated by the model. Realistic extrapolation profiles (blue dotted lines) demonstrate the generalizability of the proposed neural-ODE-based model, which could otherwise rapidly diverge because of overfitting.

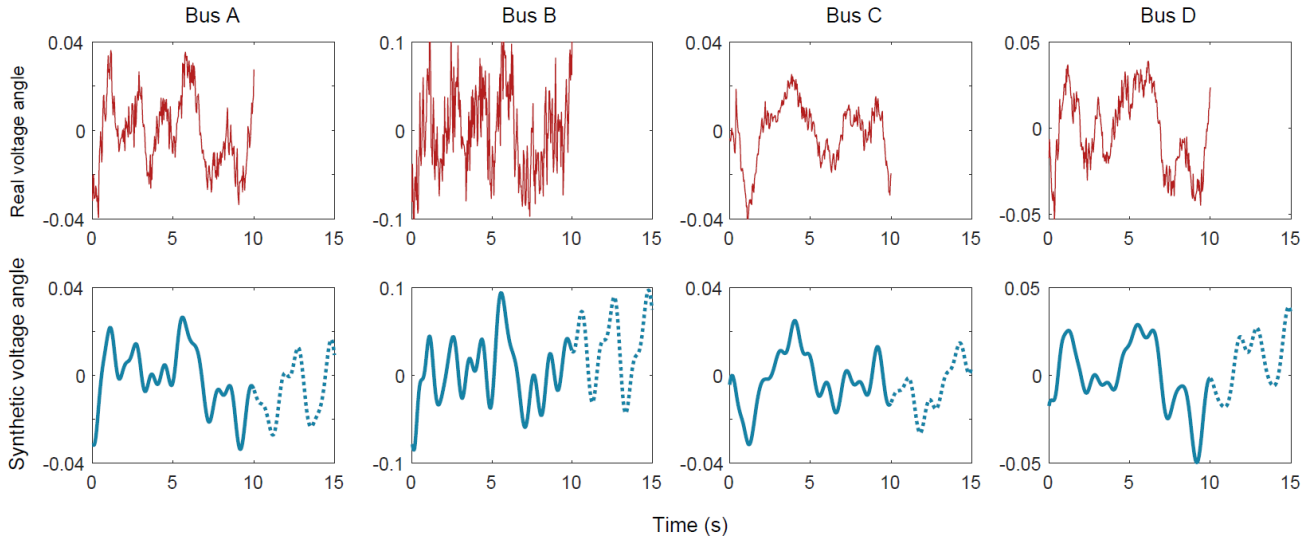

Fig. 5. Visual comparison between real and synthetic voltage angle measurements at the same selected buses for a quickly damped event that lasts for only 10 s. (a) Bus A. (b) Bus B. (c) Bus C. (d) Bus D.

We further train and test the proposed method with a 2-hour voltage oscillation event with the same settings as in Section V-B. In contrast to the quickly damped event, this weakly damped event shows more complex system dynamics, in which the voltage measurements have several changing dominant modes over time. Therefore, modal analysis is promising to validate the fidelity of the synthetic voltage oscillation measurements. To this end, the Prony method [27] is used to process both the real and synthetic voltage oscillation measurements in a moving window to analyze the dominant modes of the weakly damped oscillation over time. Here, the dominant modes refer to the modes that have relatively high energies, as specified in (6).

$$E_{\mathrm{M}} = a \sum_{i=1}^{m} \mathrm{abs}\left(\mathrm{e}^{(1/\tau + \mathrm{j}\omega)t_i}\right)^2 \tag{6}$$

where E_{M} is the energy of mode; a is the amplitude; m is the window size; ω is the mode frequency; and τ is the time constant of the mode.

Considering the large total number of modes, we select the dominant modes such that the sum of their energies account for 95% of the total energy. A synthetic time series for a certain PMU is realistic if and only if its synthetic dominant mode $\{\tau^{\mathrm{s}}, \omega^{\mathrm{s}}\}$ is close to a real one $\{\tau^{\mathrm{r}}, \omega^{\mathrm{r}}\}$.

We repeatedly perform random generation N times, as shown in Fig. 1. The fidelity rate r_i of PMU i is calculated as:

$$r_i = \sum_{j=1}^{N} I_j / N \tag{7}$$

where I_j is an indicator that shows whether the j^{th} sample is realistic according to the criteria in (8).

$$\begin{cases} \exists\{\tau^{\mathrm{r}}, \omega^{\mathrm{r}}\} \\ \text{s.t.} \left|\dfrac{\tau^{\mathrm{r}} - \tau^{\mathrm{s}}}{\tau^{\mathrm{r}}}\right| \le \% \\ \left|\dfrac{\omega^{\mathrm{r}} - \omega^{\mathrm{s}}}{\omega^{\mathrm{r}}}\right| \le \% \end{cases} \tag{8}$$

The statistics of the modal analysis of the synthetic voltage oscillation measurements for a weakly damped event that lasts for 2 hours are shown in Fig. 6, which shows the cumulative density function (CDF) of the fidelity rate of all PMUs. The fidelity rate represents the probability that the randomly created synthetic data at one certain PMU have realistic modes. Figure 6 demonstrates that the synthetic voltage oscillation data for most PMUs are realistic from the perspective of a modal analysis. We notice that the synthetic data for a small proportion of PMUs fail the modal analysis with a higher probability. This is because the corresponding PMUs are almost unaffected by the oscillation event; thus, the dominant modes correspond to random noise.

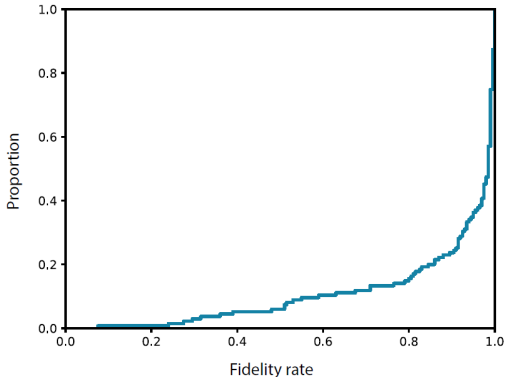

Fig. 6. Statistics of modal analysis of synthetic voltage oscillation measurements for a weakly damped event that lasts for 2 hours.

In summary, we demonstrate that the synthetic load profiles and steady-state PMU voltage measurements have realistic statistical properties and confirm that the generated voltage oscillation data have realistic oscillation modes. By combining algorithms1 and 2, we can synthetically create a massive amount of realistic eventful PMU data under generated time-varying load conditions, potentially fostering the development of data-driven methods applied to real cases.

VI. Conclusion

In this study, we propose an ML-based method to create synthetic eventful PMU data under load conditions that vary over time. Our method uses a GAN to generate load data and incorporates neural ODEs to capture the transient behavior of oscillation events that occur in a system. We utilize this method to synthetically create a massive amount of eventful PMU data under the generated time-varying load conditions and confirm that the synthetic data exhibit realistic characteristics across multiple time scales from statistical and modal analysis perspectives. The generated realistic synthetic data have the potential to alleviate the lack of real eventful PMU data and can be potentially used for the training, testing, and calibration of subsequent data-driven methods.

In general, the proposed method is feasible as long as the number of synthetic variables is less than the number of independent variables in the algebraic equations that are mainly derived from Kirchhoff's laws. Future research will extend this study to synthesize arbitrary numbers of variables with conserved algebraic relationships.

Appendix A

Table AI presents the model structure of the neural networks, where models *G* and *D* account for the generation of synthetic load profiles based on a GAN (the neural network models are implemented by TensorFlow-Keras), whereas model *f* is used to learn the voltage oscillation pattern (the neural network model is implemented by TensorFlow). MLP denotes a multiplayer perceptron followed by the number of neurons, and Conv denotes a convolutional layer followed by the number of filters. The computational environment consists of an Intel Core i7-9700 central processing unit (CPU), 32 GB of memory, and an NVIDIA RTX 2060 graphics processing unit (GPU).

TABLE AI
Model Structure of Neural Networks

Layer	*G* model	*D* model	*f* model
Input	25	900	8
Layer 1	MLP, 64	Conv	MLP, 100
Layer 2	MLP, 256	MLP, 128	MLP, 100
Layer 3	MLP, 900	MLP, 32	MLP, 4
Layer 4	Conv, 4	MLP, 1	
Layer 5	Conv, 1		

References

[1] L. Xie, Y. Chen, and P. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: linearized analysis," *IEEE Transactions on Power Systems*, vol. 29, no. 6, pp. 2784-2794, Apr. 2014.

[2] R. E. Helou, D. Kalathil, and L. Xie. (2020, Aug.). Fully decentralized reinforcement learning-based control of photovoltaics in distribution grids for joint provision of real and reactive power. [Online]. Available: http://arxiv.org/abs/2008.1231

[3] D. Wu, X. Zheng, D. Kalathil *et al*., "Nested reinforcement learning based control for protective relays in power distribution systems," in *Proceedings of 2019 IEEE 58th Conference on Decision and Control (CDC)*, Nice, France, Dec. 2019, pp. 1925-1930.

[4] T. Huang, N. M. Freris, P. Kumar *et al*., "A synchrophasor data-driven method for forced oscillation localization under resonance conditions," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 3927-3939, Mar. 2020.

[5] A. B. Birchfield, T. Xu, K. M. Gegner *et al*., "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Transactions on power systems*, vol. 32, no. 4, pp. 3258-3265, Oct. 2016.

[6] Y. Xu, N. Myhrvold, D. Sivam *et al*., "US test system with high spatial and temporal resolution for renewable integration studies," in *Proceedings of 2020 IEEE Power & Energy Society General Meeting*, Montreal, Canada, Aug. 2020, pp. 1-5.

[7] Breakthrough Energy Sciences. (2021, Aug.). A 2030 United States macro grid: Unlocking geographical diversity to accomplish clean energy goals. [Online]. Available: https://science. breakthroughenergy. org/publications/MacroGridReport.pdf

[8] D. Wu, X. Zheng, Y. Xu *et al*. (2021, Apr.). An open-source model for simulation and corrective measure assessment of the 2021 texas power outage. [Online]. Available: https://arxiv.org/abs/2104.04146v1

[9] A. Pinceti, L. Sankar, and O. Kosut. (2021, Jul).Generation of synthetic multi-resolution time series load data. [Online]. Available: https://arxiv.org/abs/2107.03547v1

[10] A. Pinceti, L. Sankar, and O. Kosut. (2021, Jul). Synthetic time-series load data via conditional generative adversarial networks. [Online]. Available: https://arxiv.org/abs/2107.03545

[11] Y. Chen, Y. Wang, D. Kirschen *et al*., "Model-free renewable scenario generation using generative adversarial networks," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3265-3275, Jan. 2018.

[12] X. Zheng, B. Wang, and L. Xie, "Synthetic dynamic PMU data generation: a generative adversarial network approach," in *Proceedings of 2019 International Conference on Smart Grid Synchronized Measurements and Analytics (SGSMA)*, College Station, USA, May 2019, pp. 1-6.

[13] X. Zheng, B. Wang, D. Kalathil *et al*., "Generative adversarial networks-based synthetic PMU data creation for improved event classification," *IEEE Open Access Journal of Power and Energy*, vol. 8, pp. 68-76, Feb. 2021.

[14] X. Zheng, N. Xu, L. Trinh *et al*. (2021, Oct.). PSML: a multi-scale time-series dataset for machine learning in decarbonized energy grids. [Online]. Available: https://arxiv.org/abs/2110.06324

[15] C. Esteban, S. L. Hyland, and G. Rätsch. (2017, Jun.). Real-valued (medical) time series generation with recurrent conditional GANs. [Online]. Available: https://arxiv.org/abs/1706.02633

[16] T. Xu, L. K. Wenliang, M. Munn *et al*. (2020, Jun.). COT-GAN: Generating sequential data via causal optimal transport. [Online]. Available: https://arxiv.org/abs/2006.08571

[17] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2019, pp. 5508-5518.

[18] Z. Lin, A. Jain, C. Wang *et al*., "Using GANs for sharing networked time series data: challenges, initial promise, and open questions," in *Proceedings of the ACM Internet Measurement Conference*, Pittsburgh, USA, Oct. 2020, pp. 464-483.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al*., "Generative adversarial nets," in *Proceedings of the 28th International Conference on Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2014, pp. 2672-2680.

[20] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang. (2017, Mar.). MidiNet: a convolutional generative adversarial network for symbolic-domain music generation. [Online]. Available: https://arxiv.org/abs/1703.10847

[21] L. Yu, W. Zhang, J. Wang *et al*., "SeqGAN: sequence generative adversarial nets with policy gradient," in *Proceedings of Thirty-first AAAI Conference on Artificial Intelligence*, San Francisco, USA, Feb. 2017, pp. 2852-2858.

[22] R. Fu, J. Chen, S. Zeng *et al*. (2019, Apr.). Time series simulation by conditional generative adversarial net. [Online]. Available: https://arxiv.org/abs/1904.11419v1

[23] M. Mirza and S. Osindero. (2014, Nov.). Conditional generative adver-

sarial nets. [Online]. Available: https://arxiv.org/abs/1411.1784
[24] R. Chen, Y. Rubanova, J. Bettencourt *et al*., "Neural ordinary differential equations," in *Proceedings of the 32nd International Conference on Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2018, pp. 6571-6583.
[25] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932-3937, Apr. 2016.
[26] B. L. Thayer, Z. Mao, Y. Liu *et al*., "Easy SimAuto (ESA): a python package that simplifies interacting with PowerWorld simulator," *Journal of Open Source Software*, vol. 5, no. 50, p. 2289, Jun. 2020.
[27] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of fluid mechanics*, vol. 656, pp. 5-28, Aug. 2010.

Xiangtian Zheng received the B.E. degree in electrical engineering from Tsinghua University, Beijing, China, in 2017. He is currently pursuing a Ph.D. degree in electrical engineering at Texas A&M Univerisity, College Station, USA. His industry experience includes an internship with PJM, Valley Forge, USA, in 2019, and an internship with Mitsubishi Electric Research Laboratory, Cambridge, USA, in 2021. His research interests include domain knowledge-informed machine learning for power system security.

Andrea Pinceti received the B. E. degree in electrical engineering from Polytechnic University of Turin, Turin, Italy, in 2015, the M.S. degree from the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, USA, in 2019, and the Ph.D. degree from the School of Electrical, Computer, and Energy Engineering, Arizona State University, in 2021. His research interests include cyber-security and data analytics related to power systems.

Lalitha Sankar received the B. Tech. degree from the Indian Institute of Technology, Bombay, India, the M.S. degree from the University of Maryland, College Park, USA, and the Ph. D. degree from Rutgers University, New Brunswick, USA. She is currently an Associate Professor in the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, USA. She currently leads an National Science Foundation Harnessing the Data Revolution (NSF HDR) Institute on data science for electric grid operations. Her research interests include applying information theory and data science to study reliable, responsible, and privacy-protected machine learning as well as cyber security and resilience in critical infrastructure networks.

Le Xie received the B. E. degree in electrical engineering from Tsinghua University, Beijing, China, in 2004, the M.S. degree in engineering sciences from Harvard University, Cambridge, USA, in 2005, and the Ph. D. degree from Carnegie Mellon University, Pittsburgh, USA, in 2009. He is currently a Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, USA. His research interests include modeling and control of large-scale complex systems, smart grids application with renewable energy resources, and electricity markets.