Large-Scale Generation and Validation of Synthetic PMU Data

Ikponmwosa Idehen, *Student Member, IEEE*, Wonhyeok Jang, *Member, IEEE*, and Thomas J. Overbye, *Fellow, IEEE*

***Abstract*—In spite of the challenges associated with obtaining actual PMU measurements for research purposes and analytic methods testing, it remains crucial that experimental input data exhibits similar quality features of real measurements for proper grid assessment and planning. The objective of this paper is to generate and validate large sets of synthetic, but realistic, PMU datasets obtained from complex grid models. A study of different variability components in PMU measurements is first presented followed by the proposed steps in generating synthetic datasets. Random variations of resource inputs are used in a simulation platform to generate prior voltage data from a synthetic 2,000-bus system, followed by a data modification process to infuse further realism into the dataset. The validation process used to assess the accuracy of the generated voltage dataset utilizes a variability metric to determine the level of inherent variations in individual measurements, and further applies a dimension reduction technique to identify the extent of electrical dynamics retained in the overall synthetic dataset.**

***Index Terms*—Phasor measurement unit, signal-to-noise ratio, power system measurements, signal synthesis, principal components.**

I. Introduction

THE EMERGENCE of fast, synchronized measurements from phasor measurement units (PMUs) and other synchrophasor devices increases the prospects of high-resolution grid monitoring. Restricted access to the power grid, classified as a critical national infrastructure, is often enforced through the use of data confidentiality contracts and non-disclosure agreements (NDA) which limit exposure to power systems information. In 2017, [1] reported the existence of 2,500 PMUs on the North American Power grid. A limited number of research-grade PMUs [2], [3] however poses a great challenge to accessing phasor measurement data, and when available are often devoid of meaningful grid contexts and dynamics of interest thus making them less relevant for scientific studies. Researchers resort to the use of experimental data, obtained from system simulations to make inferences and conclusions about the real grid. However, [4] stipulates that qualitative assessments and long-range planning of the grid can only be possible when research input data belongs to the right type and possesses high fidelity. Thus, this paper addresses the development of synthetic PMU measurements realistic enough to bear similar feature resemblance with actual PMU measurements.

Synthetic measurements have been generated in several research works in the domain of data mining, software testing, and information visualization [5]–[8]. These methods make use of intelligent techniques such as genetic algorithms, ensemble-based methods, R-programming, and rely on predefined models, patterns or random number generators to create artificial data. Due to several component and human operator interactions within the electric grid, power system measurements embed underlying grid dynamics which reflect the state of the system. Hence, power system measurements are not random nor do they strictly follow any pre-defined pattern. The work of [9] proposes the use of an intelligent generative adversarial network (GAN) machine learning technique to synthesize a realistic PMU time-series measurement. It utilizes generative and discriminative models to capture system dynamics features observed in real data which are then used in producing a synthetic time-series measurement.

To a large extent, however, these methods rely significantly on real measurements. Low quality of synthetically-manufactured data arises when confidentiality issues restrict access to data containing interesting grid phenomena. Furthermore, in the event of a production of multidimensional synthetic dataset, intelligent techniques, such as [9], may become inadequate in training multiple sets of real measurements.

Synthetic grids bearing similar characteristics and statistics with actual real electric power systems were developed in [10]–[12]. These works relied on building fictional networks similar to attributes of real complex network topologies [13]–[15], with a further inclusion of electric component models to mimic actual grid dynamics behavior. Thus, test systems with comparable grid statistics could be created to represent characteristic features of actual power system operation. This paper builds on these developed synthetic grids by proposing a method for generating high quality synthetic measurements for research use especially when confidentiality still poses great challenges to accessing real grid data.

The goal is to generate synthetic data of high resolutions typical of real PMU measurements obtained from steady dynamics-driven power systems. In comparison to

Manuscript received September 18, 2019; revised January 16, 2020; accepted February 24, 2020. Date of publication March 2, 2020; date of current version August 21, 2020. This work was supported by the Power Systems Engineering Research Center under Project T-57. Paper no. TSG-01348-2019. *(Corresponding author: Ikponmwosa Idehen.)*

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: iidehen@tamu.edu; wjang777@tamu.edu; overbye@tamu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2020.2977349

prior techniques, an advantage of the proposed method is that it is independent of the use of real measurements. Another major benefit is its ability to produce large amounts of multidimensional synthetic datasets which simultaneously embed local bus behaviors and wide-area grid dynamics in the dataset [16], [17]. The contribution of this paper is twofold – the production of synthetic PMU data using a defined framework, and feature validation showing true resemblance of synthetic datasets with real measurements. The work in this paper is categorized into the following sections. Section II analyzes the variations observed in PMU-sourced measurements, and identifies specific dynamic grid interactions which introduce measurement variabilities. The framework for producing large sets of multivariate synthetic datasets is also discussed. It employs a power systems simulation platform. Section III considers realistic demand-side and supply-side input data [18] decomposed into small time units for use in the simulation. The simulation framework and data modification process, used to improve the realism of generated data measurements, are in Section IV. Sections V and VI present the generation and validation of sample synthetic measurements. The method of principal components is used to assess the retention of underlying system electrical dynamics in the dataset, while an average variability metric is used to compare signal variations in synthetic and real dataset. Finally, the conclusions of this paper are presented in Section VII.

II. Power System Variability and Methodology for Generating Realistic Data

A. Variability in PMU Measurements

The first step in generating realistic synthetic data examines the inherent variability in field PMU measurements. Considering the use of PMU-sourced voltage measurements, persistent signal variations, ${\sigma_{\Delta V_M}}^2$ is broken down into two major components [19] – a variance component due to real system dynamics, ${\sigma_{\Delta V}}^2$, and another component, ${\sigma_\eta}^2$ whose source is attributed to the presence of measurement noise.

$$ {\sigma_{\Delta V_M}}^2 = {\sigma_{\Delta V}}^2 + {\sigma_\eta}^2 \tag{1} $$

A standard deviation σ_η is determined from a noisy signal extracted from an original measurement, and $\sigma_{\Delta V}$ is obtained from the corresponding filtered, noiseless signal. Depending on the amount of anomalies associated with certain data samples, (1) can be further revised to incorporate an additional component, ${\sigma_e}^2$ to capture variations due to PMU data errors.

$$ {\sigma_{\Delta V_M}}^2 = {\sigma_{\Delta V}}^2 + {\sigma_\eta}^2 + {\sigma_e}^2 \tag{2} $$

Depending on the phasor estimation technique and analog-digital signal conversion process, ${\sigma_\eta}^2$ will incorporate effects of quantization errors, malfunctioning current and potential transformers (CTs/PTs) used for reporting power quantities being measured and switching transients. The error component, ${\sigma_e}^2$, captures issues mostly associated with the synchrophasor device and underlying network communication. These include time skew effects in phasor angle samples which could be attributed to timing irregularities such as GPS time loss and time mis-synchronization [20]–[22], and missing measurement points due to delayed or lost data packets [23]. In previous works [24]–[26], the authors discussed some of these different anomalies associated with PMU measurements, while [27] characterized the extent of noise in field PMU data. The aggregation of the different components in (2) give rise to the overall data quality issues associated with PMU measurements.

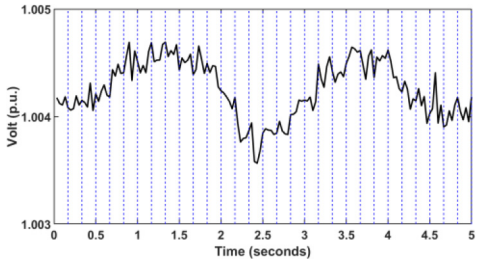


Fig. 1. 5 s per unit voltage magnitude.

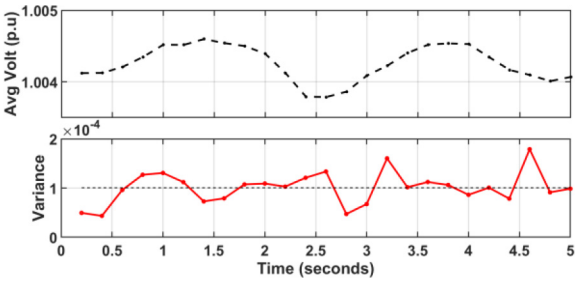


Fig. 2. Down-sampled: 5-sample window mean and variance.

Fig. 1 shows a 5-s segment of per unit (p.u.) voltage at a 345 kV bus whose time-series data is obtained from a dataset, comprising of 123 PMU measurements, sourced from different locations of a real power grid [28]. The non-stationarity feature in the time-series data in Fig. 1 is observed by the continuous, seemingly-erratic state of the individual voltage samples. A high signal quality is captured by the three decimal place representation used in uncovering the inherent variation, and the high signal-to-noise ratio (SNR) value, that is 70-dB, computed for the first minute duration of the measurement. The work of [27] shows SNR of most power system measurements to be within a range of 41 and 47-dB. Considering minimal data error, it can be concluded that the variation, $\sigma_{\Delta V_M}$ in Fig. 1 is approximately attributed to the sole contribution of $\sigma_{\Delta V}$ as in (2). An analysis of the voltage profile in Fig. 1 reveals other hidden trends by observing different down-sampled forms of the signal. Fig. 2 shows average voltage (black, dashed line) and corresponding variance (red, solid line) in consecutive, non-overlapping windows each consisting of five data samples.

A moving-window average smoothens the original profile and exposes the underlying trend, while the variance reveals the extent of fluctuation among samples. An average variance of 10^{-4} per unit, regarded as an *average variability*, is also detected when a steady change in average voltage values is observed across the non-overlapping windows. Given that σ_η and σ_e are negligible, it can be concluded that this value of

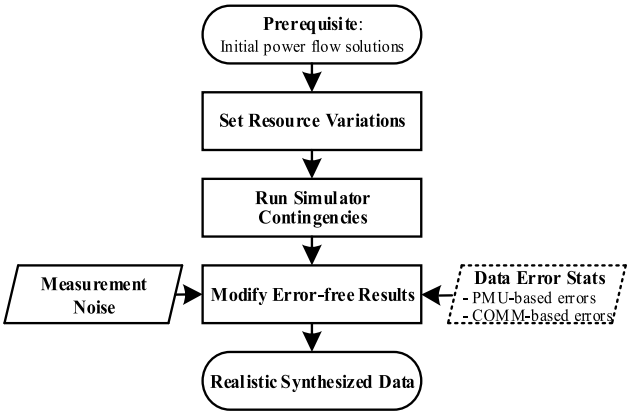


Fig. 3. Steps for synthetic data creation.

average variability provides a measure for the true fluctuation due to actual system dynamics interaction.

B. Power System Operations

Several dynamics, belonging to a wide range of time scales, are known to occur during normal grid operations [29], which in turn affects the features of PMU measurements. Generally, short time-scale events translate to small-duration transients in measurement samples, while low frequency and common-mode system variations are associated with longer time-scale events. The resulting effects are outlier samples in measurements attributed to small time-scale events such as switching events, and inherent oscillatory contents due to system loading behaviors or exciter control actions.

The effects of the different time-scale dynamics on measurements can be studied using simulations. Considering PMU measurements are often associated with timescales of one-half or one-quarter of a cycle, it is logical that simulations of similar time frames are used for the study. Here, dynamics associated with generator governor, load frequency control and boiler are preset in generator and load models already implemented in a synthetic grid. Shunts and generator reactive powers are used to conduct the voltage control aspect, while a solver in the simulator computes all system states at predefined time steps of the simulation. Availability status (ON/OFF) of generators and corresponding automatic generator control (AGC) settings are left unchanged so that selected generators and total system loads are the only control variables.

C. Methodology

Given a simulation platform, Fig. 3 shows the proposed procedure for generating synthetic measurements. It accounts for the presence of persistent signal perturbations, while simultaneously retaining underlying systems dynamics.

The methodology comprises of three main steps. Initial steps are performed on a simulation platform, followed by data modification activities on the generated, error-free simulation measurements in order to reflect actual attributes of PMU-sourced data. A prior requirement is the existence of initial power flow solution for the network case, while ensuring the absence of issues related to grid instability.

In the first step, realistic resource variations are set on both demand-side (e.g., bus loads) and supply-side (e.g., generators) locations in the system at different times. The results are multiple series of time-varying load or generation values at all system buses. Here, the aim is to incorporate the effects of $\sigma_{\Delta V}$, as shown in (2), when a grid transient simulation is run and grid states are evaluated at the different time steps. A second step requires an actual simulation of the system, with or without any selected contingency. An inclusion of a contingency, such as a line or generator outage induces additional dynamics in the error-free simulation measurements to aid the validation process. Further data modification, and an optional choice of integrating different PMU data or network-based communication errors [24] can finally be carried out to add more realism to the generated artificial dataset.

III. Time-Varying Input Resource Data Modeling

A. Time-Varying Bus Loads

The method of modeling time-varying time decomposition of resource inputs for use in a typical grid simulation introduces changing system dynamics, however in small time resolutions. The effect of this process is to partly set the value of ${\sigma_{\Delta V}}^2$ desired of the system.

Depending on the load type, variations in power system loads can follow different patterns [30]–[33], such that any load demand profile is determined by customer behaviors monitored over a duration of time. Given any preferred time resolution, any load curve can be broken into loads within equal, but smaller time resolutions (e.g., per second) to generate granular series of load values. Given loads at two consecutive time periods as L_1 and L_2, a load level L_i at any i^{th} second after L_1 is computed by interpolating between L_1 and L_2. That is,

$$L_i = L_1 + \ i \times \frac{(L_2 - L_1)}{F_d} \tag{3}$$

Here, F_d is a denominator factor used to ensure L_i is a per-second load value. For example, if the load curve is defined as an hourly or per minute interval, the values of F_d is chosen as 3600 or 60, respectively. Deductively, if the load curve is defined in 5-minute time interval, F_d will have a value of 300 (i.e., 5×60 seconds). Assuming a choice of constant load power factor, (3) can be separately implemented for both active (MW) and reactive (Mvar) load components. The inclusion of a load variation factor further aids an actualization of real load perturbations. The works of [34], [35] suggest the addition of white-noise components for the realization of practical system load variation.

$$L_{i,rdm} = (1 + \sigma_L)L_i \tag{4}$$

$L_{i,rdm}$ is an improved load obtained from a load variation factor, σ_L.

B. Time-Varying Power Generations

A system load-generation balance is ensured by setting grid generators to track changing grid load pattern provided sufficient generation exists. This is often achieved by fast-acting governor and automatic generator control systems which together contribute to variations in generator output.

Assuming similar hourly generator output variations are set to track load curves, an application of (3) and (4) can also be used to obtain small time resolution outputs. That is, between two consecutive time periods with generation levels P_1 and P_2, an i^{th} second output, P_i is given as

$$P_i = P_1 + i \times \frac{(P_2 - P_1)}{F_d} \tag{5}$$

Similarly, F_d is chosen based on the time-scale used for the consecutive time periods. Due to the predominantly high variability levels experienced by renewable sources, only nodes with attached wind generators have been considered for MW-output variations in this work.

The works of [36], [37] break down wind power variation, $P_W(t)$ into three components - a slow moving average (P_a), zero- mean fluctuating part (P_t) and a ramp event (P_r). That is,

$$P_W(t) = P_a(t) + P_t(t) + P_r(t) \tag{6}$$

The presence of prevailing wind conditions, such as wind die-out, rise, lull or gusts give rise to significant P_r-values [36]. P_a is indicative of a per-second trend power output and can be obtained from (5). P_t, attributed to wind speed turbulence is a fluctuating characteristic which can be modeled as a noise component, similar to the load discussed in Section III-A. In steady operation, P_r component is negligible and can be assumed zero. Replacing all L-load denotations in (4) with P, and introducing a σ_P-parameter to capture the associated deviation due to wind generator noise, an improved, instantaneous generation at any i^{th} second can be computed as,

$$P_{i,rdm} = (1 + \sigma_P)P_i \tag{7}$$

The value of $P_{i,rdm}$ is limited to a maximum generator output. When wind generators are used, output power is further constrained by cut-in (minimum) and cut-out (maximum) speeds [38], [39], which in turn define the extents of individual generator power production.

IV. Simulation Environment and Modification of Error-Free Simulation Data

A. Simulation Platform

The steps defined in the methodology can be performed on any power grid model. In this paper, a 2k bus synthetic power grid, shown in Fig. 4, covering the geographical footprint of Texas has been used [40].

As observed, the system comprises four transmission level voltages, and shows a predominant high concentration of wind generation in west Texas. Using the 2k-bus system, transient stability simulations are run on a grid software which emulates the operations of a typical power system. An initial case of the system provides an initial satisfactory power flow results for further simulations to be performed.

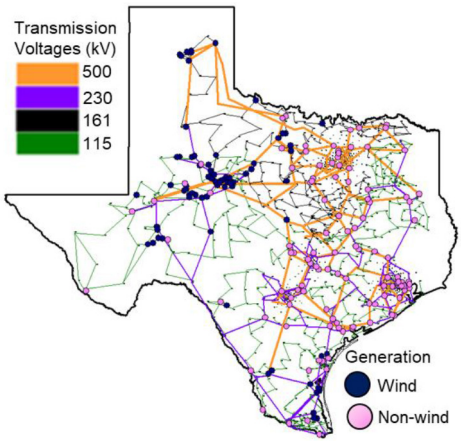


Fig. 4. Transmission voltage levels and generations in the synthetic 2,000-bus network.

With an exception to the initial case, continuous transient stability (TS) simulation is chosen over steady state, power flow evaluation carried out only at defined time steps. A benefit of using a TS-based simulation is the simulator's ability to retain complete features of all underlying system transients and dynamics. In addition to any contingency, random, time-varying bus loads and generator power outputs, generated via (3) – (7) are set as individual contingency entries in the simulator TS options. Periodic intervals, during which load and generator values are observed to vary, are also set in the contingency entry list. For this simulation, we have set periodic variations to be 5 s and 7 s for loads and generators, respectively. A simulation time step of one quarter cycles, with power flow result storage every 8 time steps, is set in a solver specification, thus mimicking PMU report rates of 30 samples per second. Upon successful completion of a 90 s simulation, the desired power flow results of voltage magnitude, angle or any other quantity can be extracted for further modifications.

The simulation of a practical grid scenario considers only a proportion of the total measurements in the 2,000-bus synthetic grid. Ninety-nine measurements have been selected according to the criteria below.

- All source measurement locations span all eight predefined geographical areas of the grid.
- All measurements cover the spectrum of the different nominal grid voltages.
- As much as possible, measurements from each area has been distributed among all nominal voltages available in that region.

B. Re-Creation of Simulation Data Output

In this paper, the process used to transform a set of initial, simulated bus voltage samples, $\boldsymbol{S} = \{s_1, s_2, s_3, \ldots\}$ into smoother variations, $\boldsymbol{S'} = \{s_{1'}, s_{2'}, s_{3'}, \ldots\}$, while introducing measurement variability, is shown in Fig. 5.

The method uses a moving-window technique, in combination with a filtering method, to smoothen out discontinuities in measurement samples. A data sample (s_i) in a non-overlapping

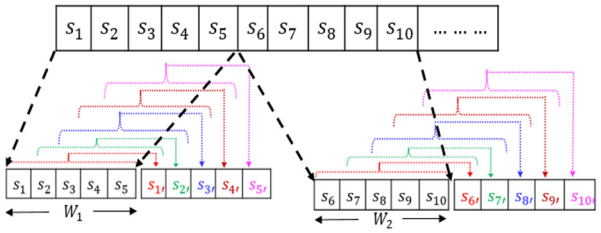


Fig. 5. Introducing variability to simulated measurements.

TABLE I
AVERAGE SYSTEM VARIABILITIES IN REAL AND SIMULATED VOLTAGE MAGNITUDE DATA

	Voltage magnitude	Voltage angle
Real	6×10^{-5}	1.5×10^{-2}
Simulation	1.1×10^{-5}	1.5×10^{-3}

data window, (W_j such that $j = 1, 2, \ldots$) extracted from the original measurement, is replaced by a new sample ($s_{i'}$), that is the sum of its window mean (s_{ave}) and a fluctuating component to a variation factor, F_v.

$$s_{i'} = s_{ave} \times (1 + F_v) \tag{8}$$

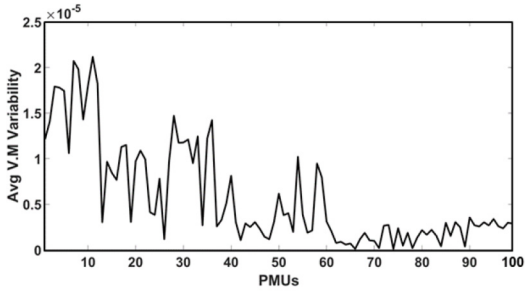
This technique of utilizing a moving-averaging window through different data segments ensures the retention of underlying dynamics, while systemic white-noise variabilities are introduced in the measurements. The effectiveness of the data re-creation process which infuses continuity into predominantly flat, disjointed, and error-free simulated voltage profiles is evaluated in the subsequent simulation result section.

V. SIMULATION RESULTS

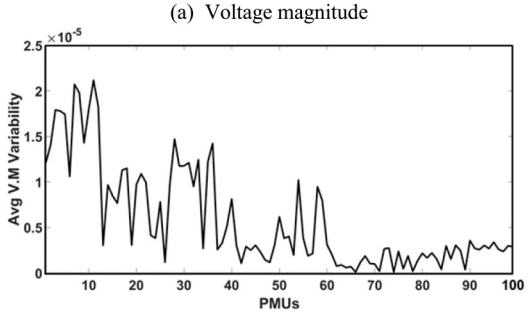
This section presents results obtained from executing the procedures in Fig. 3 for generating synthetic measurements. All analytic methods for producing resource input variations and data modification have been implemented in MATLAB, and error-free simulation results acquired by performing TS studies on a power system simulator. The synthetic measurements of interest are voltage magnitudes and angles. Respective values of σ_L and σ_P, applied system-wide at all resource input locations have been chosen based on the choice of 15-dB and 20-dB variations at loads and generation, respectively.

The extent of system variation as observed in all ninety-nine error-free, original, simulated voltage magnitude (V.M) and angle (V.A) are shown in Fig. 6 (a) and (b), respectively. They illustrate the individual average variability for all ninety-nine selected PMUs. As shown earlier in Fig. 2, the variability value of a PMU is determined by computing an average of the variances obtained over every consecutive 5-sample points in the generated time series measurements.

Here, voltage angles from all locations are observed to exhibit wider variations than magnitude. This is attributed to the use of per unit quantities of voltage magnitude, which tend to be more stable than the degree measure used for voltage angles. In addition, the presence of varying wind generations in west, south and north central synthetic Texas grid introduces higher variabilities as observed among the first 60 PMUs.



(a) Voltage magnitude



(b) Voltage angle

Fig. 6. Average variabilities in simulated measurements.

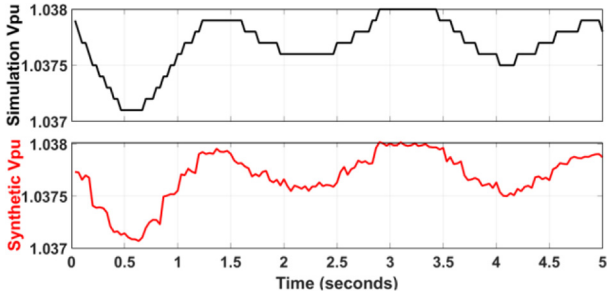


Fig. 7. Per unit voltage data: simulation versus re-created, synthetic data.

Table I shows average system measurement variabilities computed for the simulation in contrast with the 123-PMU real system. A system average voltage variability of 1.1×10^{-5} (with a maximum of 2.1×10^{-5}) is observed for the voltage magnitude when PMUS within the non-wind generation areas are ignored. In comparison with the real system, computed variabilities for both voltage magnitude and angle quantities indicate lower levels of system activities in the simulation. To mimic a higher level of dynamics, the experimental data is modified by scaling the simulation average variabilities to values similar to the real system. Selected variation factor values (F_v) of 3.0 and 10 were applied to the average voltage magnitude and average angle variabilities, respectively while the synthetic data re-creation process in Section IV-B was implemented in a 5-data sample window. Fig. 7 shows 5 s duration samples of the original, error-free, simulation per unit voltage (in black plot), alongside the re-created synthetic per unit voltage data (in red plot). The measurements have been obtained from one of the PMU locations in the system.

The discrete levels in the simulated measurement samples are due to the PMU reporting time of 0.033 s during which a value is held constant. The report interval corresponds to

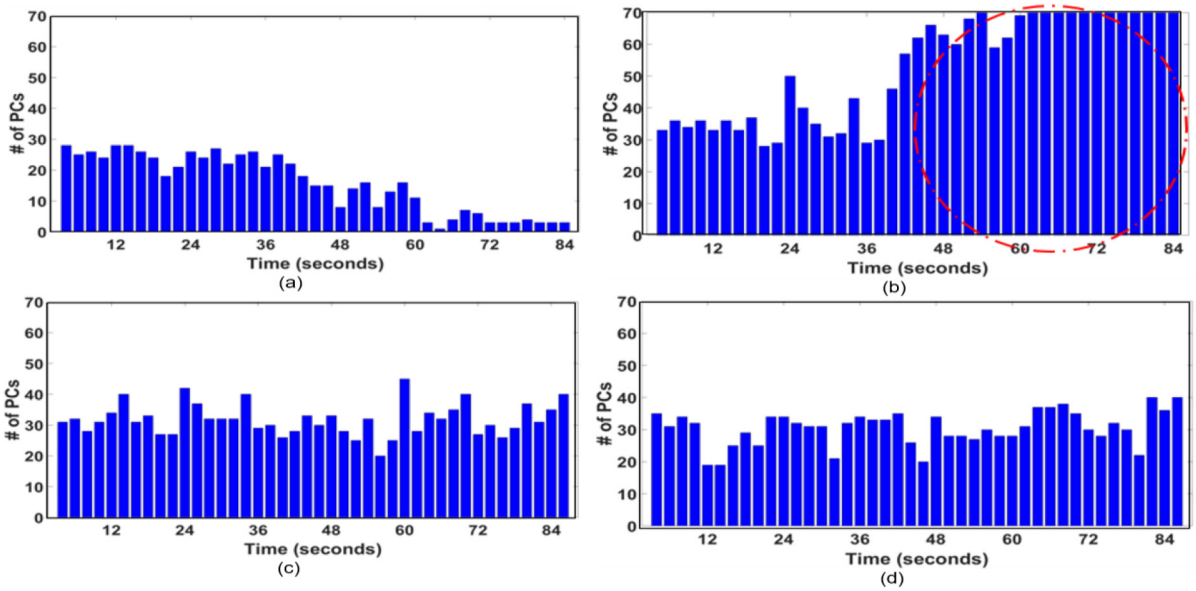


Fig. 8. *m*-counts of significant principal components in consecutive windows in (a) original, error-free simulation; (b) synthetic (partial dynamics); (c) synthetic (full dynamics) and; (d) real dataset.

eight individual step sizes, each of 4.2 ms, when system states are evaluated. Longer durations of constant values, with no perturbation, are attributed to the 'quiet' nature of grid activity. Upon transformation to the synthetic measurement, the proposed moving-average scheme is able to ensure smooth transitions between consecutive voltage samples, amidst an introduction of typical white-noise disturbances observed in real systems. Here, F_v is used to increase non-stationarity in the voltage magnitude signal during the re-creation process by distributing an average variability across samples in the original, error-free data. Depending on the need for higher signal activity and noise inclusions [41], larger values of F_v can then be used in the re-creation process.

VI. Synthetic Data Validation

This section investigates methods for validating the accuracy of the generated synthetic measurements. We consider two categories for validation - the ability of the synthetic dataset for the system to retain the underlying, electrical behavior or dynamics inherent in the original TS-simulation dataset, and a comparable average variability level with real dataset. For this purpose, principal component analysis (PCA) has been utilized.

A. Electrical Dynamics Behavior—Using PCA

The operational dynamics of any system activity can be uncovered by observing measurements obtained from devices across the grid. As a result, it is expected that the large amount of generated synthetic datasets retain an underlying grid behavior.

Given a multidimensional dataset $\boldsymbol{X}$, comprising of several bus measurements, a PCA technique [42], [43] can be used to extract key system dynamics. This is achieved by re-representing the dataset in lower dimensions, while retaining all primary attributes of the data. The technique re-expresses $\boldsymbol{X}$ consisting of n measurement locations into its most meaningful set of bases. An orthonormal matrix $\boldsymbol{P}$, known as its basis, is used to diagonalize a covariance matrix, $\boldsymbol{S_Y}$ $(= \frac{1}{n-1}\boldsymbol{YY}^T)$, such that $\boldsymbol{Y} = \boldsymbol{P}^T\boldsymbol{X}$. An eigen-decomposition of $\boldsymbol{S_Y}$ yields ordered sets of eigenvalues and eigenvectors, that is,

$$\begin{aligned} &\lambda = \{\lambda_1, \lambda_2, \ldots \lambda_m\};\ \ \lambda_1 \geq \lambda_2 \cdots \geq \lambda_m \\ &\boldsymbol{P} = \{\boldsymbol{PC_1}, \boldsymbol{PC_2}, \ldots \boldsymbol{PC_m}\};\ \ \boldsymbol{PC_i} \in \boldsymbol{R^n} \end{aligned} \tag{9}$$

m is the number of retained principal component vectors, $\boldsymbol{PC_i}$ in $\boldsymbol{P}$. The order of importance of the vectors, given as $\boldsymbol{PC_1} > \boldsymbol{PC_2} > \ldots \boldsymbol{PC_m}$, is then based on the decreasing magnitudes of eigenvalues, λ_i. Furthermore, given a threshold, percentage variance, $Variance_{th}$, which captures the major dynamics, we can compute an m number of principal components whose percentage cumulative variance, $Variance_{cumm}$ closely approximates to $Variance_{th}$.

$$Variance_{cumm} = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \times 100 = Variance_{th} \tag{10}$$

We utilize the above steps in this work to estimate the number of significant principal components in shifting time-windows, while comparing them with the original simulation dataset and those from a real system for the same duration. Fig. 8 shows the counts of significant principal components per data window in the simulation, synthetic and real voltage datasets. Here, a $Variance_{th}$ value of 98%, and data window of 4 s (or 120 measurement samples) have been used. In order to observe the effects of data modification in a portion of dataset devoid of true system dynamics, periodic resource input of generation and load are restricted to the first 50 s of the simulation.

Some of the following observations can be made from Fig. 8:

- Fig. 8(a): Given local effects of voltage, the number of components during the time period of significant system load and generation changes is observed to be

relatively large (i.e., ∼ 20-25 components) as a result of the multiple unique grid dynamics occurring at different locations of the grid. This is followed by the gradual reduction, and finally total absence of resource input variations, during which decreasing grid dynamics results in much fewer components (i.e., approximately 10 components at 50 s to one component at the end of the simulation). During this period, system dynamics is observed to be uniform or stable across the grid, whose major electrical property can then be re-represented in a very small dimensional space.

- Fig. 8(b): A data re-creation process infuses an increased level of variability individually in the now, more-realistic, synthetic data. The overall system dynamics is boosted by the variation factor used in (8), which also causes a surge in local behaviors, and hence the increased number of components in the initial time windows (25∼40 components). When compared with Fig. 8(a), a similar relative trend in the window-to-window component count in the synthetic datasets is also observed. Hence, an underlying electrical behavior of the simulation system is retained in the synthetic measurement in the presence of the newly-introduced increased activity levels.
- In Fig. 8(b), a significant observation during the period of non- inclusion of resource variation (i.e., beyond ∼ 50 s), is the number of component counts in the simulation and synthetic datasets. Large counts (58∼72 components) enclosed by the red, dashed circle are due solely to randomly-generated, fictitious data samples in the measurements at different bus locations. In contrast, when resource variations are allowed to persist beyond 50 s, a similar and moderate count of principal components is observed all through the duration of the simulation as observed in Fig. 8(c). Without a common-mode, underlying system dynamics, PCA technique detects several unique grid characteristics thus re-representing system in a dimension higher than that of the original simulation. This additional information invalidates the content of the portion of data without real system dynamics interaction.
- Fig. 8(d): Here, a typical component count (25∼40) observed in the portion of the real dataset, without any major system event occurring, is also used to validate the choice of F_v used to introduce data sample variability in the data re-creation process. This is confirmed by the similar component counts observed in the synthetic dataset of Fig. 8 (b).

B. Real System Behavior—Using Average Variabilities

In addition to satisfying the requirements of electrical behavior retention, the realism of the generated synthetic datasets is also checked by assessing their conformance to average variabilities expected of real datasets. Fig. 9 is the average voltage variability computed for each of the 123 measurements in the real dataset. Since no data anomaly has been included in any of the measurements in the synthetic dataset, we consider the use of a clean portion of the real dataset significant data anomaly across all 123 measurements. Here,

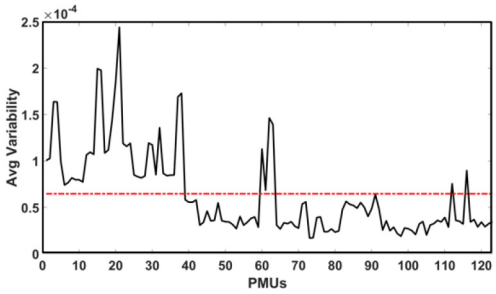


Fig. 9. Average voltage variability in the real dataset.

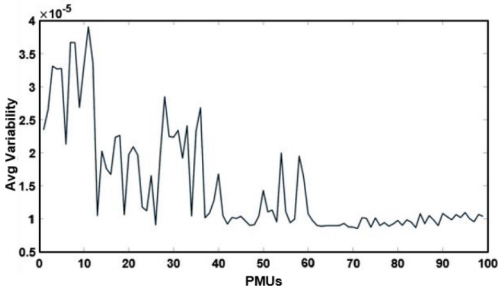


Fig. 10. Average voltage variability in the synthetic dataset.

variability in the real dataset is assessed by observing only a 5 s segment duration. A computed system average variability of 6×10^{-5} shown by the red, dotted line in Fig. 9 was used as the reference value for the synthetic data generation process. Fig. 10 is the average variability across all PMUs in the synthetic data after scaling voltage magnitude variability in the original simulation by 3.0, and introducing system variations.

In the figure, average PMU variabilities in all PMUs of the synthetic dataset are observed to fall within the same order as the system average of the real case in Fig. 9.

In terms of SNR, the black plot in Fig. 11 shows the synthetic system to possess high quality signals comparable with the real system in Fig. 12, however with higher levels of non-uniformity across the PMU measurements. Similar to the reasons mentioned in Section V, we attribute these variations to the higher activity levels of wind generation located close to the first 60 selected PMUs in the 2,000-bus grid.

In the event where a uniform signal SNRs is the preferred requirement without a corresponding change in the level of generation activity, this can be achieved by including additional variations to the measurements in the form of white noise. The red plot in Fig. 11 shows the SNR when σ_η is set to 0.0001, and incorporated in all signals. In comparison to the black plot in Fig. 11, an improvement of the overall system SNR profile is observed. PMUs located external to the vicinity of generation are shown to exhibit more uniform SNR values (80-dB). Here, increased variability was used to attain more consistent quality levels of the wide-area signals, however at the cost of system information loss in the synthetic dataset attributed to the uncoordinated inclusion of noise component. This is illustrated by the window principal component counts in Fig. 13. The uniqueness of each measurement, due

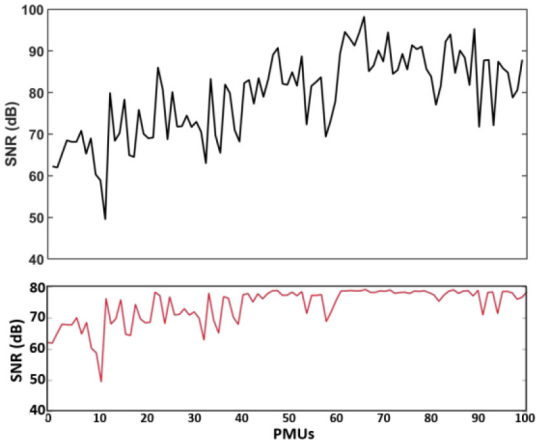


Fig. 11. Signal-to-noise ratio in the synthetic dataset.

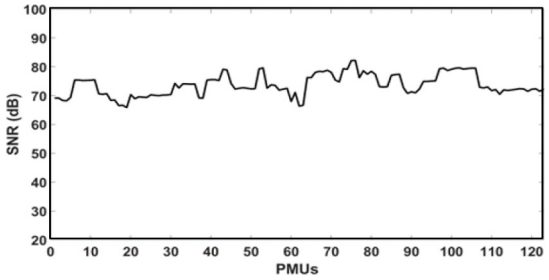


Fig. 12. Signal-to-noise ratio in the real dataset.

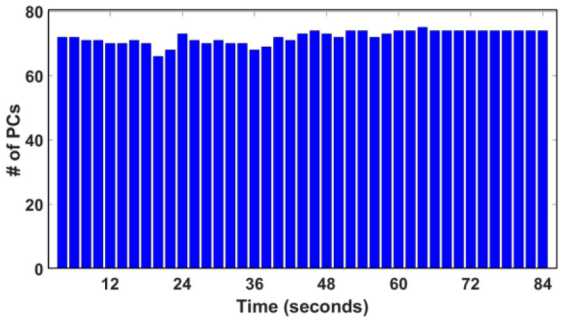


Fig. 13. Principal component counts with increased variation.

to increased levels of random signal variations, causes large counts of key principal components to be identified in each time window. True system dynamics attributed to resource input variations which were set during the initial time period of the simulation (i.e., less than 50 s) are thus masked out. A false electrical representation will assume all data windows to have similar system activity levels, hence invalidating the feature of electrical property retention.

VII. Conclusion

This paper presents a simulation framework to generate and validate realistic synthetic datasets fit for research purpose. The goal is to produce large amounts of realistic datasets from synthetic power grids, while validating the properties of the generated data. The proposed method outlines the use of realistic input resource variations and any choice of contingency within a framework for the simulation of a typical grid operation. A further modification of the resulting dataset is then performed to improve the realism of each measurement through the incorporation of measurement variations, and optional data error injections. Results show that the extent of non-stationarity and continuity features infused in the synthetic measurement are similar to real PMU data. In addition, the steps for validating large datasets from a simulation on a test power grid and synthetic data re-creation show the ability of the dataset to replicate features of real datasets. These include the capabilities to retain electrical features in the original, simulation data, consistent measurement average variation across all system measurements, and the comparable principal component counts as observed in the real dataset.

Future work will utilize statistics obtained from real datasets and other contingent events, such as loss of generators or transmission lines to generate synthetic measurements of longer time durations. Also, statistical information on time issues associated with field devices, and data loss due to the network will be leveraged to introduce data errors attributed to communication and PMU device issues. A prior knowledge of features in these synthetic datasets, and their distribution, will enable power system researchers in the testing and verification of developed data-based analysis methods.

References

[1] A. Silverstein. *Synchrophasors and the Grid*. [Online]. Available: https://www.naspi.org/sites/default/files/reference_documents/naspi_naruc_silverstein_20170714.pdf

[2] U.S. Department of Energy. (2014). *Factors Affecting PMU Installation Costs*. [Online]. Available: https://www.smartgrid.gov/files/PMU-cost-study-final-10162014_1.pdf

[3] P. Overholt, D. Ortiz, and A. Silverstein, "Synchrophasor technology and the DOE: Exciting opportunities lie ahead in development and deployment," *IEEE Power Energy Mag.*, vol. 13, no. 5, pp. 14–17, Sep./Oct. 2015.

[4] National Academies of Sciences, Engineering, and Medicine, *Analytic Research Foundations for the Next-Generation Electric Grid*. Washington, DC, USA: Nat. Acad. Press, 2016.

[5] M. Mann, O. P. Sangwan, P. Tomar, and S. Singh, "Automatic goal-oriented test data generation using a genetic algorithm and simulated annealing," in *Proc. 6th Int. Conf. Cloud Syst. Big Data Eng. (Confluence)*, 2016, pp. 83–87.

[6] G. Albuquerque, T. Lowe, and M. Magnor, "Synthetic generation of high-dimensional datasets," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2317–2324, Dec. 2011.

[7] T. Rabl, M. Danisch, M. Frank, S. Schindler, and H.-A. Jacobsen, "Just can't get enough: Synthesizing big data," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2015, pp. 1457–1462.

[8] Y. P. D. S. Brito *et al.*, "A prototype application to generate synthetic datasets for information visualization evaluations," in *Proc. 22nd Int. Conf. Inf. Visual. (IV)*, 2018, pp. 153–158.

[9] X. Zheng, B. Wang, and L. Xie, "Synthetic dynamic PMU data generation: A generative adversarial network approach," in *Proc. Int. Conf. Smart Grid Syn. Meas. Anal. (SGSMA)*, 2019, pp. 1–6.

[10] T. Xu, A. B. Birchfield, K. S. Shetye, and T. J. Overbye, "Creation of synthetic electric grid models for transient stability studies," in *Proc. 10th Bulk Power Syst. Dyn. Control Symp. (IREP)*, 2017, pp. 1–6.

[11] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3258–3265, Jul. 2017.

[12] A. B. Birchfield, T. Xu, and T. J. Overbye, "Power flow convergence and reactive power planning in the creation of large synthetic grids," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6667–6674, Nov. 2018.

[13] G. A. Pagani and M. Aiello, "The power grid as a complex network: A survey," *Physica A Stat. Mech. Appl.*, vol. 392, no. 11, pp. 2688–2700, 2013.

[14] E. Cotilla-Sanchez, P. D. H. Hines, C. Barrows, and S. Blumsack, "Comparing the topological and electrical structure of the North American electric power infrastructure," *IEEE Syst. J.*, vol. 6, no. 4, pp. 616–626, Dec. 2012.
[15] Z. Qiong and J. W. Bialek, "Approximate model of European interconnected system as a benchmark system to study effects of cross-border trades," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 782–788, May 2005.
[16] D. R. Jeske *et al.*, "Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2005, pp. 756–762.
[17] M. Wu and L. Xie, "Online detection of low-quality synchrophasor measurements: A data-driven approach," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2817–2827, Jul. 2017.
[18] T. Xu, H. Li, A. B. Birchfield, and T. J. Overbye, "Synthesize phasor measurement unit data using large-scale electric network models," 2019. [Online]. Available: arXiv:1909.03187.
[19] G. Ghanavati, P. D. H. Hines, and T. I. Lakoba, "Identifying useful statistical indicators of proximity to instability in stochastic power systems," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1360–1368, Mar. 2016.
[20] Q. F. Zhang and V. M. Venkatasubramanian, "Synchrophasor time skew: Formulation, detection and correction," in *Proc. North Amer. Power Symp. (NAPS)*, 2014, pp. 1–6.
[21] Q. Zhang, V. Vittal, G. Heydt, Y. Chakhchoukh, N. Logic, and S. Sturgill, "The time skew problem in PMU measurements," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2012, pp. 1–6.
[22] C. Huang *et al.*, "Data quality issues for synchrophasor applications—Part I: A review," *J. Mod. Power Syst. Clean Energy*, vol. 4, no. 3, pp. 342–352, 2016.
[23] L. E. Miller, A. Silverstein, and D. Anand, "PMU data quality: A framework for the attributes of PMU data quality and a methodology for examining data quality impacts to synchrophasor applications," North Amer. SynchroPhasor Initiative, Richland, WA, USA, Rep. PNNL 26313/NASPI-2017-TR-002, 2017.
[24] I. Idehen, W. Jang, and T. Overbye, "PMU data feature considerations for realistic, synthetic data generation," 2019. [Online]. Available: arXiv:1908.05244.
[25] I. Idehen, Z. Mao, and T. Overbye, "An emulation environment for prototyping PMU data errors," in *Proc. North Amer. Power Symp. (NAPS)*, 2016, pp. 1–6.
[26] I. Idehen and T. J. Overbye, "PMU time error detection using second-order phase angle derivative measurements," in *Proc. IEEE Texas Power Energy Conf. (TPEC)*, 2019, pp. 1–6.
[27] M. Brown, M. Biswal, S. Brahma, S. J. Ranade, and H. Cao, "Characterizing and quantifying noise in PMU data," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, 2016, pp. 1–5.
[28] PJM. (Nov. 8, 2018). *Private Generic PMU Data*. [Online]. Available: https://www.pjm.com/markets-and-operations/advanced-tech-pilots/private-generic-pmu-data.aspx
[29] P. W. Sauer and M. A. Pai, *Power System Dynamics and Stability*. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.
[30] D. K. Ranaweera, G. G. Karady, and R. G. Farmer, "Economic impact analysis of load forecasting," *IEEE Trans. Power Syst.*, vol. 12, no. 3, pp. 1388–1392, Aug. 1997.
[31] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," *IEEE Trans. Power Syst.*, vol. 16, no. 3, pp. 498–505, Aug. 2001.
[32] PSMT Department. (2014). *Fundamentals of Transmission Operations*. [Online]. Available: https://www.pjm.com/-/media/training/nerc-certifications/trans-exam-materials/foto/foto-lesson4-load-forecasting-and-weather.ashx?la=en
[33] U.S. Department of Energy. (2016). *Maintaining Reliability in the Modern Power System*. [Online]. Available: https://www.energy.gov/sites/prod/files/2017/01/f34/Maintaining%20Reliability%20in%20the%20Modern%20Power%20System.pdf
[34] Y. Qiu, J. Zhao, and H. D. Chiang, "Effects of the stochastic load model on power system voltage stability based on bifurcation theory," in *Proc. IEEE/PES Transm. Distrib. Conf. Expo.*, 2008, pp. 1–6.
[35] V. S. Perić and L. Vanfretti, "Power-system ambient-mode estimation considering spectral load properties," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1133–1143, May 2014.
[36] H. Banakar, C. Luo, and B. T. Ooi, "Impacts of wind power minute-to-minute variations on power system operation," *IEEE Trans. Power Syst.*, vol. 23, no. 1, pp. 150–160, Feb. 2008.
[37] E. A. DeMeo, W. Grant, M. R. Milligan, and M. J. Schuerger, "Wind plant integration [wind power plants]," *IEEE Power Energy Mag.*, vol. 3, no. 6, pp. 38–46, Nov./Dec. 2005.
[38] E. Muljadi and C. P. Butterfield, "Pitch-controlled variable-speed wind turbine generation," *IEEE Trans. Ind. Appl.*, vol. 37, no. 1, pp. 240–246, Jan./Feb. 2001.
[39] R. Karki, H. Po, and R. Billinton, "A simplified wind power generation model for reliability evaluation," *IEEE Trans. Energy Convers.*, vol. 21, no. 2, pp. 533–540, Jun. 2006.
[40] *Electric Grid Test Case Repository*. Accessed: Mar. 15, 2019. [Online]. Available: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/
[41] S. Wang, J. Zhao, Z. Huang, and R. Diao, "Assessing Gaussian assumption of PMU measurement error using field data," *IEEE Trans. Power Del.*, vol. 33, no. 6, pp. 3233–3236, Dec. 2018.
[42] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
[43] J. Shlens, "A tutorial on principal component analysis," 2014. [Online]. Available: arXiv:1404.1100.

Ikponmwosa Idehen (Student Member, IEEE) received the B.Eng. degree from the University of Benin, Benin, Nigeria, in 2008, the M.S. degree from Tuskegee University, AL, USA, in 2014, and the Ph.D. degree from Texas A&M University, College Station, TX, USA, in 2019, where he is currently a Postdoctoral Researcher with the Department of Electrical and Computer Engineering.

Wonhyeok Jang (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Sungkyunkwan University, Suwon, South Korea, in 2008 and 2010, respectively, and the Ph.D. degree from the University of Illinois at Urbana–Champaign, IL, USA, in 2017. He is currently a Lecturer with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.

Thomas J. Overbye (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 1983, 1988, and 1991, respectively. He is currently a TEES Eminent Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.