Validation of Synthetic U.S. Electric Power Distribution System Data Sets

Venkat Krishnan, *Senior Member, IEEE*, Bruce Bugbee, Tarek Elgindy, Carlos Mateo, Pablo Duenas, Fernando Postigo, Jean-Sébastien Lacroix, Tomás Gómez San Roman, *Senior Member, IEEE*, and Bryan Palmintier, *Senior Member, IEEE*

***Abstract*—There is a strong need for synthetic yet realistic distribution system test data sets that are as diverse, large, and complex to solve as real systems. Such data sets can facilitate the development of advanced algorithms and the assessment of emerging distributed energy resources while avoiding the need to acquire proprietary critical infrastructure or private data. Such synthetic data sets, however, are useful only if they are realistic enough to look and behave similarly to actual systems. This paper presents a comprehensive framework for validating synthetic distribution data sets using a three-pronged statistical, operational, and expert validation approach. It also presents a set of statistical and operational metric targets for achieving realistic data sets based on detailed characterization of more than 10,000 real U.S. utility feeders. The paper demonstrates the use of the proposed validation approach to validate three large-scale synthetic data sets developed by the authors representing Santa Fe, New Mexico; Greensboro, North Carolina; and the San Francisco Bay Area, California.**

***Index Terms*—Electric distribution test feeders, synthetic data sets, validation, statistical metrics, power flow, smart grid use case.**

I. Introduction

DISTRIBUTION system test data sets that represent real utility systems are key to the development of cutting-edge, scalable power system algorithms and to conducting realistic estimates of grid interactions with distributed energy resources (DERs). Very few data sets from utilities are available in open access ([1]–[3] represent a few exceptions), and these data sets represent only very small subsets of realistic utility service territories, typically a few isolated representative feeders. As such, working with real data typically requires nondisclosure agreements, and hence there are restrictions on how the research findings get disseminated. Openly available test feeders are alternative options; however, the focus of these has been on the evaluation of distribution system analysis techniques and tools rather than on accurate representations of real systems [4]. Hence, they do not capture the complexity, size, or diversity of typical utility systems. For advanced use cases, it is desirable to have large systems that serve millions of diverse customers through multiple interconnected feeders and substations; controlled by voltage-regulating devices; and tied together by reconfigurable switching devices, breakers, and reclosers. Additionally, component- and voltage-level diversity and geographic coordinates that enable linking relevant weather and renewable data are also highly desirable [5].

Therefore, there have been recent efforts by the research community to create synthetic yet realistic distribution data sets that are as complex, diverse, and large as real systems. A subset of authors of this publication have developed algorithms for creating synthetic European and U.S. Reference Network Model (RNM) distribution data sets [6]–[8]. The focus of this paper is on the methods, metrics, and results from validating such synthetic data sets and on quantitatively ascertaining their realism by comparing them with real utility data sets.

As described in detail in Section II, although there have been some recent contributions of approaches for validating synthetic transmission systems [9]–[13] and some past work in validating synthetic European-style distribution feeders [14]–[17], there are notable gaps in the comprehensive validation of U.S.-style distribution systems. This paper fills this gap by proposing a three-pronged validation approach for electric distribution system data sets and applying it to validate first-of-their-kind, very large-scale (up to 10 million electric nodes) synthetic U.S. distribution system data sets. Past published work (see Section II) creating test distribution data sets used either power flow feasibility or statistical summaries of a few selected system size or graph metrics to develop

Manuscript received September 11, 2019; revised January 15, 2020; accepted February 21, 2020. Date of publication March 20, 2020; date of current version August 21, 2020. This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract DE-AC36-08GO28308. This work was supported by the Advanced Research Projects Agency-Energy (ARPA?E) under the GRID DATA Program. A portion of this research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) and located at the National Renewable Energy Laboratory. Paper no. TSG-01341-2019. *(Corresponding author: Venkat Krishnan.)*

Venkat Krishnan, Tarek Elgindy, and Bryan Palmintier are with the Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, CO 80401 USA (e-mail: venkat.krishnan@nrel.gov).

Bruce Bugbee was with the Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, CO 80401 USA. He is now with Oracle, Redwood City, CA 94065 USA.

Carlos Mateo, Fernando Postigo, and Tomás Gómez San Roman are with the Institute for Research in Technology, Comillas Pontifical University, 28015 Madrid, Spain.

Pablo Duenas is with the MIT Energy Initiative, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Jean-Sébastien Lacroix was with CYME, Eaton Corporation, St-Bruno-de-Montarville, QC J3V 3P8, Canada. He is now with BBA, Mont-Saint-Hilaire, QC J3H 6C3, Canada.

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.2020.2981077

TABLE I
VALIDATION CRITERIA AND IMPORTANCE

Criteria	Importance
Realistic physical layout	Stakeholder acceptance, weather, demographics
Realistic system size	Scalability beyond one feeder, reconfiguration options, multi-feeder interactions
Realistic topology and components ratings	System performance, standard equipment ratings, critical foundation for all use cases
Realistic reconfiguration options	Automated reconfiguration, post-reconfiguration operations simulation
Comprehensive load specification	Basic power flow, enable rich scenarios, time-series load, customer and load types
Representative power flows and voltage profiles	Realistic scenarios, capturing key concerns for distribution operations
Computational complexity	Typical solution times, challenging scenarios

the validation process. Here, however, we bring these aspects together—including a broad list of metrics and their recommended target values from real systems—into one formal process. Specific contributions that support this three-pronged approach include.

a) *Statistical Validation:* For the first prong, we present a statistical categorization and grading process that uses histograms of system metrics. The process answers the question, "Could the sample of metric statistics from the synthetic system come from real system metric distributions?"
b) *Metrics:* In support of the statistical validation, we propose a comprehensive set of statistical and operational metrics that can be used to validate the test systems per the criteria listed in Table I. These ensure that the metrics are as comprehensive as possible, covering various factors, including physical layout, realistic size, graph structure, and operations. Additionally, the paper provides a target distribution for these metrics by reporting the aggregated distributions from several real-world utility data sets (>10,000 real-world feeders for some metrics).
c) *Operational Validation:* For the second prong, we propose using commercial and open-source power flow software to assess additional metrics, such as power flow convergence, loss levels, and American National Standards Institute (ANSI) limit nonviolations in planning load levels.
d) *Expert Validation:* For the third prong, we propose actively involving industry experts to add a vital dimension of engineering judgment to the validation process. This helps to capture real-world design features and broaden applicability of the data sets beyond academic researchers to include industry users and vendors.
e) *Quantifying Improvements Over the Current State of the Art:* This paper also presents comparisons of currently available open-access test feeders against target metric distributions (or validation regions) estimated from aggregated utility data (Fig. 20 in appendix). Such discussions reveal their smaller sizes and lack of equipment diversity, including the scarcity of low voltage (LV) secondary and meshed urban system data.

The paper is organized as follows. Section II provides a comprehensive background on past work for electrical test system validation. Section III summarizes the synthetic data set creation process for the U.S.-style distribution data sets validated in this paper. Section IV introduces the proposed three-pronged validation process, involving *statistical*, *operational*, and *expert* validation steps. Section V presents numerical illustrations validating three large-scale synthetic data sets. Section VI presents the conclusions. An appendix provides supporting validation materials, including histograms of metrics from U.S. utility systems related to equipment size and design. The metrics from several utilities are aggregated to ensure that utility-specific data are anonymized and yet easily disseminated for widespread use in the validation processes.

II. BACKGROUND: TEST DATA SET VALIDATION

For transmission model data, [9] used component rating data as the criteria for validation. It was found that existing public test cases lacked key details of transmission thermal limits and generator capability curves, and therefore [9] proposed a data-driven approach to estimate the missing parameters from other publicly available data sets. Reference [10] used system size, topology, and component ratings as the criteria to validate synthetic networks representing the central Illinois (200-bus ACTIVSg200) and western South Carolina (500-bus ACTIVSg500) systems. The baseline metrics of system proportions (e.g., number of buses per substation, load levels, generation capacities) and component ratings (e.g., transformer and transmission line parameters, capacities) were extracted from the 70,000-bus Eastern Interconnection, 16,000-bus Western Interconnection, and North American grid data sets from the Federal Energy Regulatory Commission Form 715. Reference [11] expanded to add graph-based metrics—such as minimum spanning tree, Delaunay triangulation, Delaunay neighbors—for validating a synthetic 2,000-bus Texas grid. Graph-based metrics of complex networks were also used by [12] to study 15 different European transmission networks at the 200-kV and 400-kV levels. Degree distribution, characteristic path length, network diameter, betweenness centrality, and global clustering coefficient were some of the metrics used to characterize the real transmission networks and later to develop synthetic European transmission grids that are topologically consistent with real European power grids [13].

Efforts to validate synthetic distribution data sets have been largely limited, especially for U.S.-style distribution grids. The Distribution System Operators (DSO) Observatory project [14] in Europe collected realistic system size, component ratings, and representative power flow solutions from DSOs and used them as the primary criteria for synthetic data set validation. The system size and performance metrics—e.g., consumers per area or per substation; ratio of LV to medium-voltage (MV) consumers; LV or MV circuit length per LV or MV supply point; underground-to-overhead line ratio at LV and MV levels; typical transformer capacities per consumer for LV, MV, and HV systems; load per consumer, power flow, and reliability metrics—were obtained from 190 large distribution system operators in Europe. Such

data have been used by the RNM to build synthetic European grids with real geospatial coordinates starting with building location, parcel use, and street maps information (European DSOs [15] and Croatian DSO [16]); however, the large differences in distribution design between Europe and the United States minimize the applicability of such tools to develop and validate U.S. systems.

Reference [17] developed a method to create validated synthetic radial-type U.S. distribution systems. The method relied on estimating the parametric fits of metric distributions (e.g., cable length, graph diameter) using real-world utility system data, and it used samples from those parametric distributions to develop synthetic distribution grids. The validation was done by comparing the metric distributions from synthetic and real grids using the Kulback-Leibler (KL) distance-based divergence index. Although statistically precise, such approaches might miss key aspects of utility systems that result from engineering-based design. The data sets and validation described in this paper fall within another body of work where the synthetic data set creation process does not directly use the metric distributions from real utility data sets because the actual metrics from a single utility might be sensitive and hamper the dissemination of results due to data protection concerns. Therefore, alternative methods [6], [7], [18] are used that build the synthetic systems (lines, transformers, and voltage control devices) from the ground up, starting with customer location, population census, and street maps, much like the aforementioned European efforts [15]. The outputs are typically geospatial topologies of feeders, with a diverse component mix. Validating such rich and detailed systems typically requires using a diverse set of metrics obtained from several real utility grids.

In this paper, we demonstrate the proposed three-pronged validation process by applying it to the validation of medium- and large-scale synthetic data sets depicting Santa Fe, New Mexico (SAF); Greensboro, North Carolina (GSO); and the San Francisco Bay Area, California (SFO). These U.S.-style data sets are publicly accessible in OpenDSS, CYME, and Geographic Information System (GIS) formats at the DR POWER [19] and BetterGrids.org [20] repositories.

III. Building Synthetic U.S. Distribution Systems: A Summary

The synthetic distribution system creation process used for the data sets to be validated in this paper builds on the use of the U.S. Reference Network Model (RNM-US) that automatically builds large, U.S.-style distribution networks from customer load and location data using equipment lists and design practices informed by actual utility data, as described in detail in [7], [8]. Fig. 1 shows a summary of various steps in the data set creation process. The process starts with the selection of an existing area for which information about all buildings' location, footprint, height, and customer type is available, typically from parcel data. Each building represents a consumption point for which the peak load is estimated in two steps: 1) identify the building use from the parcel data (e.g., single- or multifamily house, hotel, hospital, industry, school, or restaurant); and 2) estimate the peak load of each consumer through linear interpolation by using a database of building reference models (e.g., commercial buildings [21]) and by assuming the energy consumption is correlated with the building volume. The peak load also allows inferring the voltage level (high, medium, or low) at which each consumer or building connects. Street maps are obtained from public sources such as the OpenStreetMap project or databases from GIS offices of cities and counties. The cost and technical parameters of the components in the catalog (i.e., typical equipment sizes, electrical parameters, installation cost) are typically obtained from utility data and surveys. With this information, the RNM-US builds the distribution system.

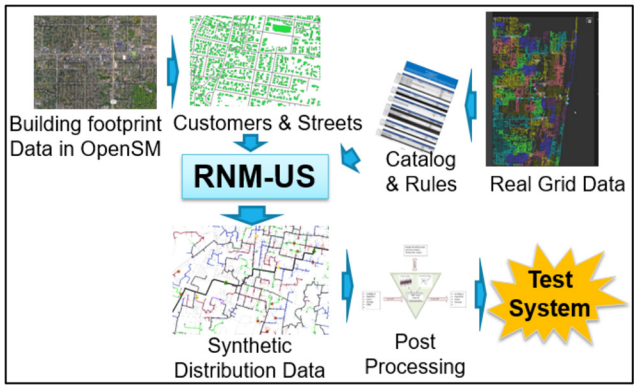


Fig. 1. Process to build the synthetic distribution data sets.

The algorithms of the RNM-US are from ground up [7] in the following sequence: locating and sizing secondary transformers (MV to LV), planning the LV lines, locating and sizing MV substations, planning the MV feeders, locating and sizing transmission substations, and, finally, planning the sub-transmission network. There are additional stages related to designing loops and installing switching devices to improve reliability (i.e., minimizing energy not served), installing voltage control devices in the MV system (capacitor banks and voltage regulators), and allocating the phase connections for customer loads to reduce unbalance [8]. The planning problem, therefore, is subject to electrical (voltage and thermal), geographic (street maps, settlements, and topography), and reliability constraints. The output from RNM-US includes the GIS files and an OpenDSS description to run power flows.

The RNM-US output is post-processed using the National Renewable Energy Laboratory's DiTTo[1] tool to add details, including the multi-transformer bank arrangement for substations, populating additional switches/reclosers at strategic locations for reliability, adding complete descriptions of control settings for utility voltage equipment (load tap changers (LTCs), regulators, and capacitors), and adjusting the design to ensure realistic voltage profiles. This step also includes converting the base data sets to multiple analysis formats (e.g., OpenDSS, CYME, and DEW) to support multi-tool

[1] https://github.com/NREL/ditto: Distribution Transformation Tool, an open-source framework that enables programmatic post-processing and editing of distribution network models as well as translations between data formats.

analysis. Next, the base system is validated as described in this paper. The validation process often initiates several iterations of the Fig. 1 process, including modifying RNM-US and changing its inputs. Once this base system is refined and validated, the final step again uses the DiTTo post-processing tool in conjunction with other tools[2] to add time-series loads and rich scenarios for DERs and control aspects.

IV. Three-Pronged Validation Method

Fig. 2 presents the developed three-pronged validation process, which includes: 1) *statistical validation* that compares synthetic data sets against real U.S. distribution system data from utilities, 2) *operational validation* that simulates use cases of interest and extracts operational metrics for comparison against typical real-world behavior, and 3) *expert validation* that assesses the qualitative design aspects that might have been missed while encouraging industry and end user buy-in.

A. Statistical Validation

Statistical validation consists of three steps.

1) Calculating Metrics for Both Real and Synthetic Data: The following metrics are used, mapped to Table I criteria.

a) Physical layout and topology: Average degree and distance measures such as graph diameter and characteristic path length of a feeder [22]. These metrics are defined as follows. The degree (k_i) of node i in a graph with n total nodes and internode adjacency matrix A (with elements a_{ij}) is defined by (1), which computes the number of lines at a node in a feeder. The average degree, therefore, will estimate the average number of lines connected at a feeder node. The graph diameter (d_{max}) is given by (2), where d_{ij} denotes the minimum number of links needed to traverse to get from node i to node j, which can be estimated from the off-diagonal elements of the adjacency matrix A. The largest of all such internode shortest distances will estimate the graph diameter of the feeder system. Finally, the characteristic path length (L) is the average of all internode shortest paths, d_{ij}, in a feeder.

$$k_i = \sum_{j=1}^{n} a_{ij} \tag{1}$$

$$d_{max} = \max_{ij,\, i \neq j} d_{ij} \tag{2}$$

$$L = \frac{1}{n(n-1)} \sum_{\substack{\forall i,j \\ i \neq j}} d_{ij}. \tag{3}$$

b) Realistic physical size: Number of connected customers, substations, feeders, and transformer capacity.

c) Realistic electrical design and equipment parameters: At LV, MV, and high-voltage (HV) levels, electrical design can be quantified in terms of line lengths—for example, single-phase line length, single- and two-phase line length (especially at the MV level), and three-phase line length. They can be further divided into proportions of overhead (OH) and underground (UG) lines. Density metrics can be a ratio of total line length (or transformer size) to customers or square miles.

d) Reconfiguration options: Number of substation breakers, reclosers, sectionalizers, switches, and lateral fuses.

e) Load specification: Active and reactive power demand.

f) Voltage control schemes: Number of capacitor banks, voltage regulators, and substation tap settings.

These metrics can be further expanded to include other use cases—for instance, equipment outage probabilities and minimum time to repair for reliability-related use cases.

2) Defining Validation Regions Using Aggregated Real Data: The distributions of each metric from the aggregated utility data are classified as typical (80% of the observed data), uncommon (15% of the observed data), and rare (remaining 5%, in the tails). This is in line with the typical statistical convention, where anything outside the 95% confidence region is labeled as rare, and the choice of 80% within the normal region is to add an additional layer of stratification for typical vs. uncommon. Depending on the data type, different methods are used to estimate these regions. For continuous data, either empirical quantiles or high-probability density regions (HDR) are used to identify respective regions that cover the desired proportion of data. The HDR method is better suited for distributions that show multimodal properties (possibly resulting in disjointed ranges of data for a specific validation region). Categorical data are handled by ranking the prevalence of each group and finding the appropriate coverage.

3) Grading Synthetic Metric Distributions: If the metric distributions from utility data are for the same region as the synthetic data set, comparison of the distribution functions can be performed using KL-divergence-type distance measures [17]. But there are issues in publishing a wide range of statistics from a specific utility and in applying such techniques to regions managed by a different utility. To circumvent this issue and yet disseminate the anonymized real data for research reproducibility, this paper proposes using aggregated metrics from several regional utilities and comparing the aggregated metric distribution with the synthetic data set metric for a specific region. Therefore, the goal of the statistical validation is not to exactly match these two distributions but to make sure the synthetic data capture the variety in metrics found in reality while highlighting any

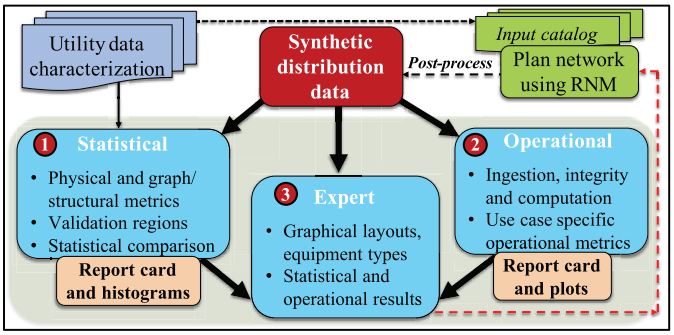


Fig. 2. Three-pronged validation process: an overview.

[2]https://www.nrel.gov/grid/smart-ds.html: ***R2PD*** (Resource to Power Data) is a power system modeler-friendly tool for populating wind and solar power time series and forecast data for T&D data sets with geographic coordinates.

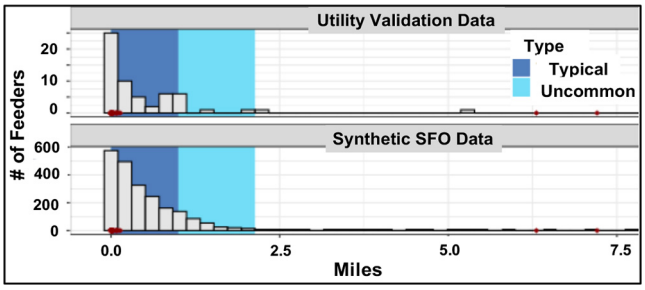


Fig. 3. Total LV 3ϕ line length per feeder: utility (58 feeders) vs. SFO data set.

values notably outside of the aggregated utility distribution (indicating unrealistic values) for further assessment. In other words, the goal of the statistical validation process is to assess the validation exit criteria by asking, "*Is it feasible that the metrics from the synthetic data sets could arise from a real system*?" To answer this, we compare the densities of metrics from synthetic feeders against validation regions estimated using aggregated utility data, and we assign grades of good/marginal/check based on the rare region density. If more than 10% of the synthetic feeders have a metric in the rare region, we consider that metric to deviate too much from the normal observation and give it a "check" grade (meaning further iterations of data set creation might be needed). Anything between 0%–5% and 5%–10% will be deemed "good" and "marginal," respectively, thereby allowing for diversity within normal observations.

As an illustration of the statistical validation process, Fig. 3 shows the total LV three-phase line length metric per feeder for real utility data sets (top plot) and the SFO synthetic data set (bottom plot). Using the utility data, typical (between [0+, 1] mile) and uncommon (between [1, 2.14] miles) regions have been identified. The y-axis shows the feeder counts. Upon comparison, about 85% of synthetic SFO feeders fall in the typical region, 12.4% in the uncommon region, and 2.6% in the rare region. Therefore, based on the rare region density, this metric is deemed "good." The final validation report card will contain the results of such comparisons for many other metrics (see Table IV). Fig. 3 also shows red markers along the x-axis, which are for the existing open test feeders [4] (Fig. 20 y-axis in the appendix). Note that only a few open test feeders provide LV secondary data, with limited diversity.

B. Operational Validation

Operational validation uses standard distribution system analysis tools to check system integrity (e.g., connectivity, default values for missing parameters, computation time) during data ingestion and operational responses for selected use cases based on power flow and short-circuit analysis. Table II shows the operational metrics proposed for validation, which are given a single target value or range based on the input of industry experts familiar with the analysis of countless utility systems across North America.

This approach also avoids the computationally prohibitive need to simulate very large numbers of real-world feeders under myriad loading scenarios that would otherwise be required to develop descriptive statistics or comparative distributions of the metrics. In addition to these criteria, operational validation is used to assess feeder voltage histograms and profiles (voltage vs. distance) for the synthetic data sets to ensure realistic voltage performance and effective voltage control settings.

TABLE II
OPERATIONAL METRICS AND VALIDATION EXIT CRITERIA

Analysis	Parameters (Unit)	Exit Target
Power flow convergence	Number of iterations	< 20
	Time to solve (s)	< 30
Power flow results	Minimum service voltage (p.u.)	> 0.95 [23]*
	Maximum service voltage (p.u.)	< 1.05 [23]*
	Average voltage (V)	0.95–1.05
	System losses (%)	< 10
	Overloads count	<0 [23]
	Undervoltage count	<0 [23]
	Overvoltage count	<0 [23]
	Voltage regulation range/bandwidth	Range
	LTCs within range/bandwidth	Range
	Transformer loading (%)	0–100
	Voltage unbalanced factor (%)	<3 [24]
Short-circuit (SC)	Transmission SC level (kA) (>138 kVLL)	20–40 [25]
	Subtransmission SC level (kA) (69 kVLL < V < 138 kVLL)	20–40 [26]
	Medium-voltage SC level (kA) (1 kVLL < V < 69 kVLL)	0.3–40 [24]
	Low-voltage SC level (kA) (<1 kVLL)	0.5–100 [24]

**For most loads (>99.5%). See discussion in Section V-B.*

C. Expert Validation

The expert validation stage solicits qualitative utility, vendor, and other stakeholder inputs based on the review of GIS system visualizations, one-line system diagrams, statistical and operational metric histograms, and validation process results. This allows performing sanity checks, capturing engineering judgments and design features (e.g., switch placement and types, LV design) that are not easily found in the utility data.

V. NUMERICAL ILLUSTRATION

Fig. 4 and Table III present the three data sets developed using the method summarized in Section III. The data sets capture diverse geography and designs: SAF models the Santa Fe, New Mexico, distribution system (with multiple feeders) covering urban and suburban regions. GSO models a larger system covering industrial, suburban, and rural customers in Greensboro, North Carolina. SFO, the largest of the three, spans a diverse mix of rural, urban, suburban, and industrial customers across 12 counties in the San Francisco Bay Area, California.

Tables V and VI in the appendix summarize the building types and estimated load sizes in the synthetic SFO data set. As shown, the SFO feeders largely serve urban and suburban-type customers, although they also include rural feeders in outlying areas (as shown in Fig. 4).

A. Statistical Validation

The first step in the statistical validation was to collect and process large quantities of real-world U.S. electric distribution

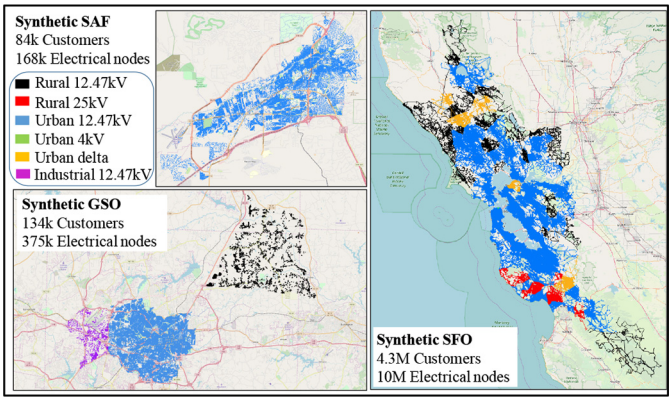


Fig. 4. SMART-DS synthetic distribution system data sets: SAF, GSO, SFO.

TABLE III
SYNTHETIC DATA SETS—SUMMARY STATISTICS

Statistics (counts) for data sets	SAF	GSO	SFO
Buildings	38,590	70,554	2,265,594
Medium voltage	31	144	1,535
Low voltage	38,558	70,407	2,264,014
Customers	84,169	134,882	4,299,805
Medium voltage	31	495	11,503
Low voltage	84,138	134,387	4,288,297
Buses	88,886	181,631	4,916,869
Electrical nodes	168,005	375,334	9,868,205
Transmission substations	2	7	148
Subtransmission substations	8	31	632
Distribution transformers	11,300	25,933	559,151
Power lines (length in km)	1,921	4,534	116,837
High voltage (subtransmission)	27	167	4,128
Medium voltage	966	2,302	64,460
Low voltage (secondary)	928	2,107	48,249
Primary feeders	28	98	2,236

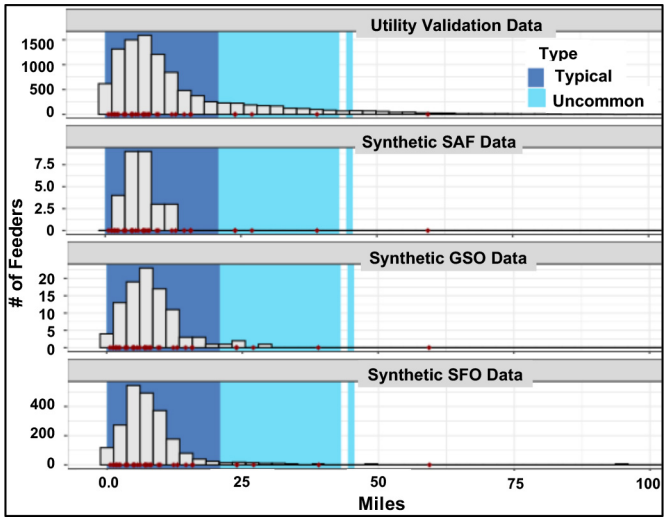


Fig. 5. Total MV 3ϕ line length per feeder: utility (10,149 feeders) vs. synthetic.

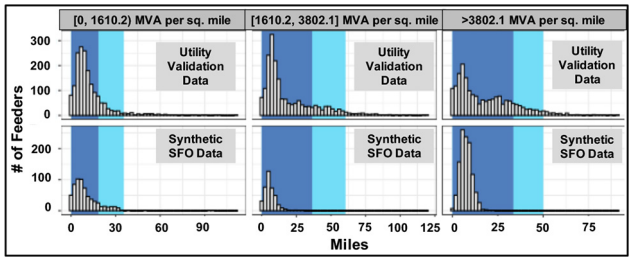


Fig. 6. Total MV 3ϕ line length per feeder (load density partitions): aggregated utilities data vs. SFO.

data. For this effort, we collected electrical models for thousands of actual feeders plus summary data for additional distribution systems from multiple U.S. utilities representing both large and small utilities and service territories across the country (see acknowledgments). These data were parsed using the DiTTo[1] to automate metric gathering for system design (e.g., line lengths, density), equipment sizes, and graph parameters. The utility data were not used directly as inputs to the synthetic system generation to ensure that the creation and validation process stayed independent and that real-world data were anonymized.

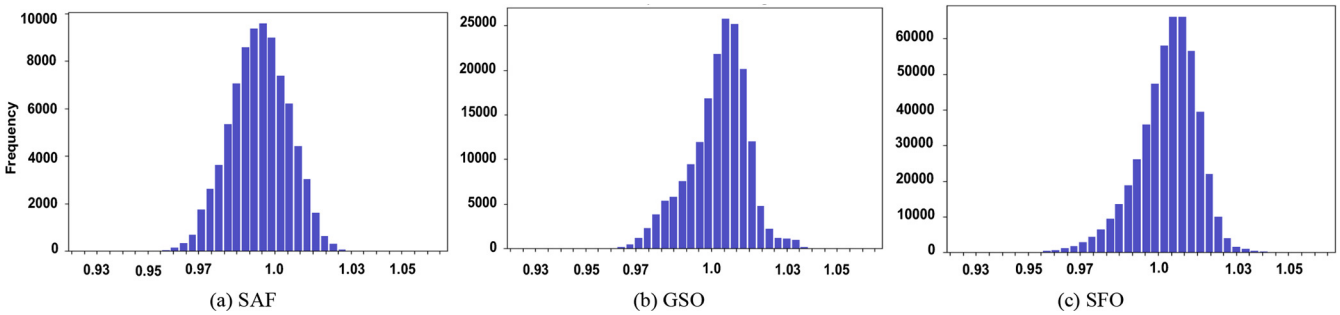
Table IV presents the statistical validation report card for the three synthetic data sets, including the validation criteria (Table I), related metrics, comparison results (good, marginal, check), validation region definition from aggregated utility data, and the number of utility feeders used in the statistical comparison for each metric. For all three data sets, nearly all metrics have passed the test, meaning we can infer that, statistically, the data in the synthetic system could have originated from real-world utility systems. Note that the goal is to attain "good" grades for as many metrics as possible (to follow time-tested grid design practices); however, not every metric needs to get "good" grades. Instead, "marginal" or "check" are indications to look further to understand any discrepancies while recognizing that some additional diversity is observed in real-world systems. To validate this concept, we compared metrics from one actual utility to those from among all utilities and found that although most metrics were "good," one metric "percentage of overhead three-phase line length" received a "check" with 67% in the rare region density, indicating different diversities among utilities.

As an example, Fig. 5 shows the set of histograms of total MV three-phase line length per feeder for aggregated utilities vs. synthetic data sets. By design, the SFO data set captures more diversity in this metric than the SAF data set. The GSO and SFO data sets have comparatively more feeders with loads connected via many three-phase MV lines (reflective of their larger size and customer base).

To further explore, the SFO results are broken down by load density in Fig. 6. The metric "*load per square mile of area*" was chosen as a proxy [27], [28] for identifying rural (<1610.198 MVA/mi^2), suburban (>1610.198, but <3802.112 MVA/mi^2), and urban (> 3802.112 MVA/mi^2) feeders based on clustering of real utility data. The distributions show diversity in every load density partitions, with more feeders in higher density urban regions of the SFO. The appendix provides more metric comparison histograms showing both the diversity and realism of the synthetic data sets.

TABLE IV
STATISTICAL VALIDATION RESULTS (PER FEEDER METRICS): SAF, GSO, AND SFO DATA SETS (G = GOOD, MGL = MARGINAL, CHK = CHECK, GRD=GRADE, TYP = TYPICAL, UNC = UNCOMMON, RAR = RARE, ϕ = PHASE, OH = OVERHEAD)

Metric (per feeder)	SAF Data Set Results				GSO Data Set Results				SFO Data Set Results				Utility Data Validation Regions		
	Typ	UnC	Rar	Grd	Typ	UnC	Rar	Grd	Typ	UnC	Rar	Grd	Typical	Uncommon	#Feeder
Dist. Xfrmr Tot. (MVA)	0.86	0.14	0.00	G	0.84	0.15	0.01	G	0.79	0.20	0.00	G	[0+, 1.73], [4.94+, 31]	[1.73+,4.94], [31+,38.629]	5923
Tot. real load (kW)	0.79	0.21	0.00	G	0.66	0.31	0.03	G	0.52	0.37	0.11	Chk	[4181+, 13793]	[577+,4181], [13793+, 17590]	1330
LV 1ϕ line len. (miles)	1.00	0.00	0.00	G	1.00	0.00	0.00	G	1.00	0.00	0.00	G	[0+, 34.75]	[34.75+, 44.31]	57
LV 3ϕ line len. (miles)	1.00	0.00	0.00	G	0.55	0.34	0.11	Chk	0.85	0.12	0.03	G	[0+, 1]	[1+, 2.135]	58
MV 1&2 ϕ line ln.(mile)	1.00	0.00	0.00	G	1.00	0.00	0.00	G	0.99	0.01	0.00	G	[0+, 35.36]	[35.36+, 124.62]	10632
MV 3ϕ line len. (miles)	1.00	0.00	0.00	G	0.96	0.04	0.00	G	0.97	0.03	0.00	G	[0+, 20.84]	[20.84+, 45.6]	10149
MV OH 1&2 ϕ line ln (mi)	0.93	0.07	0.00	G	0.97	0.03	0.00	G	0.91	0.09	0.00	G	[0+, 19.1]	[19.1+, 84.5]	10099
MV OH 3ϕ line ln (mile)	1.00	0.00	0.00	G	0.96	0.04	0.00	G	0.95	0.04	0.00	G	[0+, 17.7]	[17.7+, 39.7]	9747
% of OH 1&2 ϕ lines	0.89	0.07	0.04	G	0.82	0.18	0.00	G	0.81	0.18	0.01	G	[0+, 0.23], [0.46+, 1]	[0.23+, 0.46]	9350
% of OH 3ϕ lines	0.86	0.14	0.00	G	0.88	0.12	0.00	G	0.89	0.09	0.02	G	[0.4+, 1]	[0.18+, 0.4]	9492
# Cust.	1.00	0.00	0.00	G	0.78	0.20	0.02	G	0.76	0.20	0.04	G	[94+, 2607]	[8+, 11837]	9734
Ratio of MV 1&2 ϕ line len. to Cust.	1.00	0.00	0.00	G	0.95	0.05	0.00	G	0.87	0.12	0.01	G	[0+, 0.12]	[0.12+, 0.24]	9221
Ratio of MV 3ϕ line len. to Cust.	0.64	0.36	0.00	G	0.80	0.20	0.00	G	0.70	0.27	0.03	G	[0+, 0.09]	[0.09+, 0.77]	8556
# Fuses	1.00	0.00	0.00	G	0.85	0.15	0.00	G	0.82	0.18	0.00	G	[4+, 187]	[187+, 281]	6013
# Reclosers	0.96	0.04	0.00	G	0.97	0.03	0.00	G	0.94	0.06	0.00	G	[0+, 5]	[5+, 9]	6013
# Regulators	1.00	0.00	0.00	G	1.00	0.00	0.00	G	1.00	0.00	0.00	G	[0+, 3]	[3+, 8]	11574
Sectionalizer	1.00	0.00	0.00	G	1.00	0.00	0.00	G	1.00	0.00	0.00	G	[0+, 1]	[1+, 3]	5020
# Switches	0.82	0.14	0.04	G	0.74	0.20	0.05	Mgl	0.75	0.19	0.06	Mgl	[3+, 392]	[392+, 635]	5020
# Cap. Banks	1.00	0.00	0.00	G	1.00	0.00	0.00	G	1.00	0.00	0.00	G	[0+, 5]	[5+, 7]	11574
Avg. degree	1.00	0.00	0.00	G	0.87	0.13	0.00	G	0.89	0.09	0.01	G	[1.9+, 2.06]	[1.6+, 1.9], [2.06+, 2.1]	5020
Char. path len. (miles)	0.96	0.04	0.00	G	0.89	0.11	0.00	G	0.87	0.11	0.01	G	[12.4+, 95]	[2+, 12.4], [95+,134.39]	5020
Graph dia. (miles)	0.96	0.04	0.00	G	0.90	0.10	0.00	G	0.88	0.11	0.01	G	[32+, 260]	[4+, 32], [260+, 371]	5020



(a) SAF (b) GSO (c) SFO

Fig. 7. Histogram of nodal voltages: SAF, GSO, and SFO data sets (meeting ANSI criteria for steady-state voltages between 0.95–1.05 p.u.).

B. Operational Validation

OpenDSS (http://sourceforge.net/projects/electricdss), CYME (http://www.cyme.com), and DEW (https://www.edd-us.com/dew-ism/) were used to conduct operational validation to assess the accuracy of the topology and performance of the synthetic models. The first step of operational validation simply involved ingesting the models into the simulation software. This invoked the built-in system consistency and input connectivity checks, typically included in commercial software that offer rich tools for parameter checking and warnings. During this phase, a number of required improvements were identified and corrected in the data sets, including: 1) line segmentation and coordinate assignment for better visualization, section identification, and faster distributed simulations; 2) equipment state and connectivity refinements, including correcting unwanted loops from erroneous settings

of normally open switches; 3) line parameter corrections (e.g., line susceptance) and transformer model corrections (e.g., impedances range between 2%–30%); 4) LV configuration (North American 120/240-V center-tapped transformers, where appropriate) and substation transformer LTC setting corrections (typically at 1.05 to avoid end-of-line low voltages); and 5) protection equipment, breakers, and fuses, and ratings matching with line ratings to avoid overload warnings.

After import, engineering simulations (power flow and short-circuit analysis) were conducted for operational validation against the metrics reported in Table II. The loading level of the base power flow simulation resembled a medium-load level, and it did not result in significant overloads (many transformers were loaded around 40%). The power flows for this planning loading scenario converged in about 12 iterations in CYME. The SAF data set took about 15 s, whereas GSO took about 35 s. To manage the computational burden, the GSO data set was simulated separately for the industrial, rural, and urban regions. Similarly, the larger SFO system was divided into 40 regional systems (each with multiple substations and feeders) that were simulated separately. The system technical losses per feeder in the SFO system were around 3.71% on average, with the 25th, 50th (median), and 75th percentiles at 0.8%, 2.73%, and 4.9%, respectively. About 96% of the SFO primary feeders had losses less than the target 10% maximum (Table II). Because typical values for electrical losses can be different depending on the type of feeder (e.g., typically lesser in urban vs. higher in rural), such a statistical range is acceptable and corroborates with open literature (e.g., [29] observed a peak load loss of 4.8% on average and 5.8% as 75th percentile in Electric Power Research Institute (EPRI) green circuits). The peak short-circuit currents at various voltage levels also met the Table II criteria (the SAF and GSO systems had 34.77 kA and 34.8 kA at the transmission, 20.74 kA and 25. 91 kA at the sub-transmission, 6.487 kA and 7.748 kA at the MV and 4.921 kA and 4.748 kA at the LV levels, respectively).

A key aspect of distribution system operations, and hence the operational validation, is regulating the voltage. The histograms in Fig. 7 of nodal voltages for all the feeders in the SAF, GSO, and SFO data sets show that the vast majority of voltages are within the ANSI C.84 Range A limits of 0.95–1.05 V p.u. (Table II criteria); however, a few undervoltages were observed. Because such undervoltages do occur in practice, we opted to expand the acceptance criteria as follows: 99.5% of customers in ANSI Range A (0.95–1.05 V p.u.), >99.95% under 1.05 V p.u. (without distributed generation), and >99.999% of customers >0.94 V p.u. Where these criteria were not met, three approaches were used in post-processing to correct: 1) increase substation transformer kVA, 2) increase LTC set point to 1.05 p.u. voltage output (up from 1.03 default), and 3) where needed, add additional line regulators based on voltage profiles. After these corrections, the SAF data set had no voltage <0.94 V p.u. and about 0.005% nodes between 0.94–0.95 V p.u. The GSO data set had no undervoltage violations. The SFO data set had all voltages between 0.9287 and 1.0579 V p.u., with 99.9998% of nodes >0.94 p.u. (excluding 12 nodes of ∼10 million), 99.9917% >0.95 p.u. (excluding 412 nodes), and 99.9924% of nodes below 1.05 p.u. (excluding 377 nodes above 1.05 p.u.).

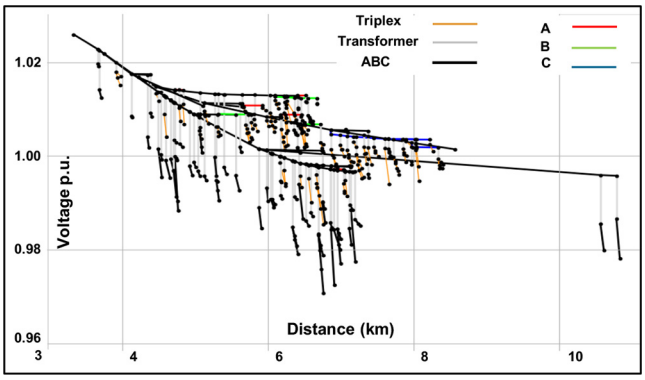


Fig. 8. Selected SFO industrial feeder: voltage vs. distance.

Fig. 8 shows the voltage vs. distance plot for a representative 12.47-kV industrial feeder in the SFO data set. The red, blue, and green show different single-phase lines. Black lines are three-phase, while orange lines refer to triplex lines (LV service connections). The thin gray lines (which appear as sudden drops) show transformers. This figure shows both the realistic behavior seen with these models and hints at the importance of including the LV triplex secondary cables (not easily found in utility data and other open-access data sets) in power flow analyses that typically result in 0.005–1.0 V p.u. voltage drop.

C. Expert Validation and Key Design Considerations

The third element of validation solicited qualitative inputs from industry experts. These discussions provided a wealth of suggestions around layout, standard practices, and other aspects difficult to capture quantitatively. Specifically, this input improved: MV and LV line lengths, number of distribution transformers per feeder (initially low), number of customers connected per distribution transformer, number of MV consumers per feeder, feeder-tied switches for system reconfiguration and reliability (initially low), loops within the same feeder (e.g., remove unwanted loops), transformer reactance in substations (e.g., making adjustments based on typical X/R ratios), location and count of capacitors/regulators, breakdown of distribution transformer size by phase (less single-phase 75-kVA types), and the graphic representation.

The LV (secondary) system configuration and single- and three-phase lines were also thoroughly assessed, and the RNM-US inputs and heuristics were adjusted to produce more realistic layouts (i.e., typical treelike secondary layouts with limited star configurations). Additionally, the substations were designed to include feeder connection grouped by transformer bank with a default of four possible connections each, including some redundant unconnected feeder heads for future growth (which improved substations-to-feeders and substations-to-total load ratios). HV/MV substations were to be designed with at least one circuit breaker/recloser located in proximity and having typical transformer configurations (commonly delta-wye and occasionally delta-delta). Utilities' equipment standardization practices were followed using

standard power ratings and voltage levels for various equipment. This information helped bound the planning algorithm dimension and achieve realistic metric distributions. Urban-suburban regions used higher fractions of underground cables (typically, underground-to-overhead line ratios are about 30% in the United States, as shown by Fig. 11 in the appendix) and were made to primarily rely on capacitors for voltage regulation, unlike rural and lower load density regions that also used voltage regulators. Capacitor banks were installed at three-phase buses, whereas fuses were added to single-phase laterals (fuse count was proportional to the single-phase line lengths, i.e., lesser fuses per feeder in high-load-density urban feeders compared to low-load-density rural feeders). More switches were installed in feeders with underground cables (high-load-density urban) than overhead lines, with some switches identified as remote-controlled for advanced reconfiguration and reliability use cases. Pad-mounted three-phase distribution transformers were massively deployed for three-phase consumers, whereas single-phase center-tap transformers served single-phase consumers.

In summary, these improvements identified during the expert validation process tremendously enhanced the realism of the synthetic systems. Such details in the system creation process are not easily found from real utility data or statistical distributions, and thereby dependence on expert feedback as the third layer in this three-pronged validation process revealed such finer details for synthetic system post-processing and improvement. Some system upgrades (e.g., adding appropriate equipment coordinates, switches at relevant locations, and system segmentation capabilities) were also made to ensure easier adoption of these very large synthetic data sets (>2000 feeders) by end users and seamless ingestion of these data sets into proven industry software, such as CYME and EDD, which were used as part of the operational validation.

VI. Conclusion

This paper presented a comprehensive process to validate synthetically produced distribution data sets and identified a comprehensive set of quantitative metrics for validation of U.S. distribution designs. The paper demonstrated the approach and results of validating three large-scale open-access synthetic data sets: versions of Santa Fe, New Mexico (SAF) (urban and suburban); Greensboro, North Carolina (GSO) (industrial, suburban, and rural); and the San Francisco Bay Area, California (SFO) (rural, urban, suburban, and industrial with additional design diversity). The proposed three-pronged validation process—consisting of statistical, operational, and expert validation—is shown to quantitatively and qualitatively ascertain the realism of synthetic systems, including comparisons with real data.

Looking ahead, this validation approach can be applied to other synthetic U.S. electric distribution systems data sets. Opportunities for additional research include enhancing the set of metrics, such as by directly incorporating the X/R ratio; developing techniques to more accurately capture measures of design diversity and the corresponding conditional distributions for key metrics; and expanding operational validation to directly capture topics of interest in grid modernization, such as hosting capacity or system performance under future scenarios.

Increasing interest in integrated transmission-and-distribution (T&D) data sets represents a further research opportunity. The overall three-pronged approach presented here should still be applicable for T&D, and the specific metrics and results presented can be used to assess the realism of the distribution portion of T&D. Other existing approaches and metrics can be used for the transmission-only portion; however, integrated T&D metrics have yet to be defined.

TABLE V
Summary of Buildings and Loads in SFO Bay Area Counties

Counties	Count *1000s*	1ph / 3ph LV %	Rural / Urban %	Load *MW*	Urban %
San Benito	326.5	98.1 / 1.9	0.5 / 99.5	1013.5	99.7
Sonoma	466.9	98.0 / 1.9	0.6 / 99.4	1968.9	99.7
Napa	49.9	99.3 / 0.7	10.6 / 89.4	100.1	87.9
Solano	83.9	98.5 / 1.5	1.5 / 98.5	263.4	99.2
San Francisco	54.9	98.6 / 1.4	4.9 / 95.1	142.0	95.7
Marin	21.4	99.1 / 0.9	12.9 / 87.1	47.5	90.8
Contra Costa	91.8	98.3 / 1.7	0.0 / 100.0	272.4	100
Alameda	132.9	98.3 / 1.7	2.5 / 97.5	419.5	98.6
Santa Cruz	251.5	97.5 / 2.5	0.2 / 99.8	916.1	99.9
San Mateo	233.5	99 / 1.00	1.8 / 98.2	593.3	98.5
Lake	128.4	98.1 / 1.9	3.8 / 96.2	506.1	98.0
Santa Clara	214.2	99 / 1.00	5.0 / 95.0	464.7	95.4

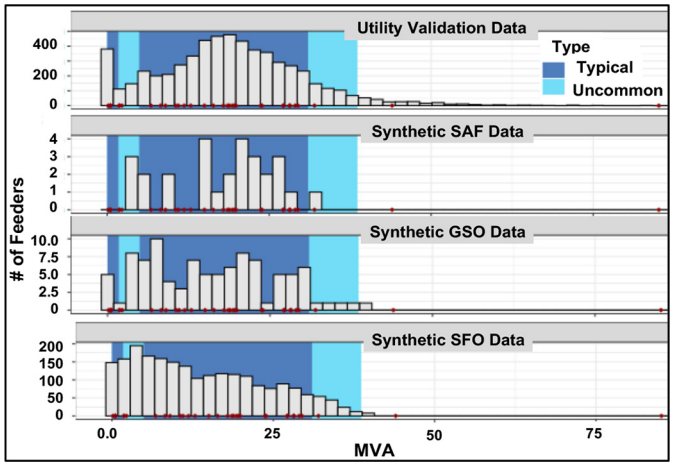


Fig. 9. Total distribution transformer capacity per feeder (5,923 real feeders).

Appendix

Table V summarizes the building counts, phase connections, and demographics served by the synthetic SFO data set. A high percentage of customers in all counties are in urban/ suburban regions, and they are connected at single-phase LV; there are fewer three-phase LV customers. The large MV customers are less than 0.01% in most counties. Table VI elaborates the customer types and their load levels. Single-family units form the bulk of single-phase LV connections, whereas supermarkets and industrial loads form the majority of three-phase loads.

Figs. 9–20 present histograms for several metric comparisons, which should be interpreted with following notes in mind.

TABLE VI
SUMMARY OF TOTAL LOAD BY CUSTOMER TYPES IN BAY AREA COUNTIES SERVED BY SYNTHETIC SFO DATA SET

Type	Single-family unit			Multi-family unit			Other 1*			Other 2**		
Counties	Total MW	Rural (%)	Urban (%)	Total MW	Rural (%)	Urban (%)	Total MW	Rural (%)	Urban (%)	Total MW	Rural (%)	Urban (%)
San Benito	839.3	0.08	82.73	105.8	0.00	10.44	18.5	0.01	1.81	49.9	0.21	4.72
Sonoma	1600.7	0.11	81.19	222.5	0.00	11.30	48.8	0.01	2.47	96.8	0.19	4.73
Napa	79.9	7.17	72.66	5.8	1.20	4.55	10.6	2.74	7.85	3.8	0.99	2.83
Solano	205	0.35	77.47	23.0	0.01	8.73	16.8	0.03	6.36	18.6	0.39	6.66
San Francisco	112.2	2.70	76.32	15.5	0.81	10.09	7.6	0.22	5.11	6.8	0.54	4.21
Marin	35.3	2.88	71.53	1.6	0.06	3.38	1.4	0.40	2.45	9.2	5.91	13.39
Contra Costa	161.6	0.00	59.32	79	0.00	28.99	11.8	0.00	4.35	20	0.00	7.34
Alameda	342.8	0.49	81.22	36	0.01	8.56	17	0.48	3.58	23.7	0.47	5.18
Santa Cruz	679	0.01	74.11	165.5	0.00	18.07	30.1	0.02	3.26	41.5	0.11	4.42
San Mateo	293.2	0.18	49.24	35.3	0.01	5.94	251.6	1.03	41.37	13.2	0.27	1.95
Lake	422.5	0.62	82.86	32.7	0.01	6.46	15.7	0.13	2.96	35.2	1.19	5.76
Santa Clara	390.5	2.07	81.95	6.4	0.26	1.12	63.2	1.74	11.87	4.6	0.52	0.48

*Other 1 = Retail + Mall + Supermarket + Hotel + Restaurant; ** Other 2 = Warehouse + Industrial + Education + Office + healthcare*

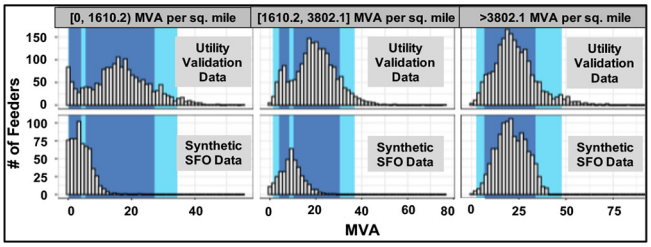


Fig. 10. Total distribution transformer MVA per feeder (by load density): aggregated utility data vs. SFO data.

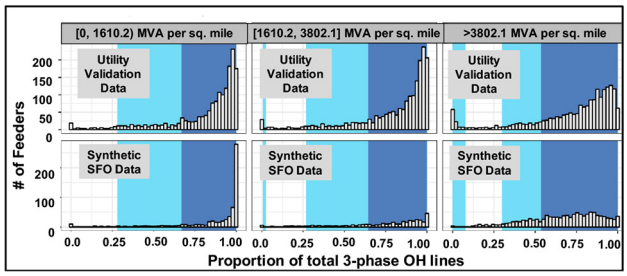


Fig. 11. Proportion of total 3-phase OH lines per feeder, 1670:1871:1887 utility feeders in each partition vs. SFO data.

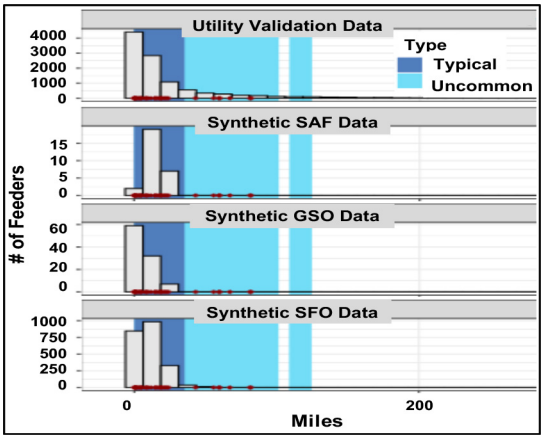


Fig. 12. Total MV 1- and 2-phase line length per feeder (10,632 utility feeders).

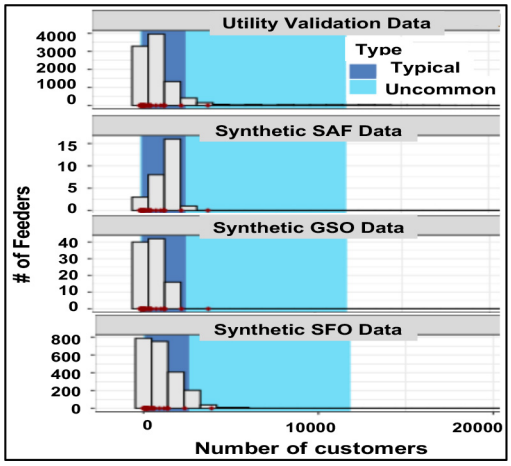


Fig. 13. Number of customers per feeder (9,734 real utility feeders).

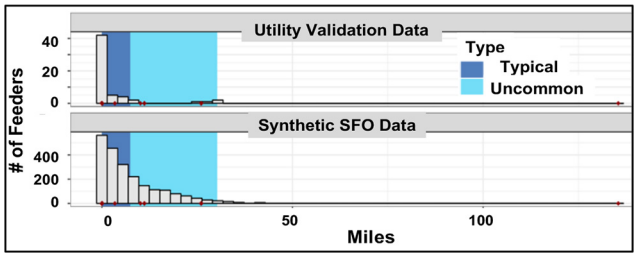


Fig. 14. Total LV 1-phase line length per feeder: utility (57 feeders) vs. SFO (2236 feeders). Note that most actual feeders have a considerable extent of LV connections for smaller customers, but these data are seldom captured in utility models, which results in the very low count of LV data in the utility validation plot above. Nevertheless, the synthetic SFO system fills that gap and has a range of total LV length values consistent with the validation region.

- The histograms compare aggregated metric distributions from several utilities to a synthetic data set for a region. In line with the goals of the statistical validation explained in Section IV-A3, all the synthetic metric distributions are a subset of the aggregated utility distributions.
- The synthetic metrics capture diversity while reflecting the customer base being served. For example, Fig. 10 clearly shows how the high-load-density urban feeders (right subplot) have a closer match between the synthetic SFO and aggregated

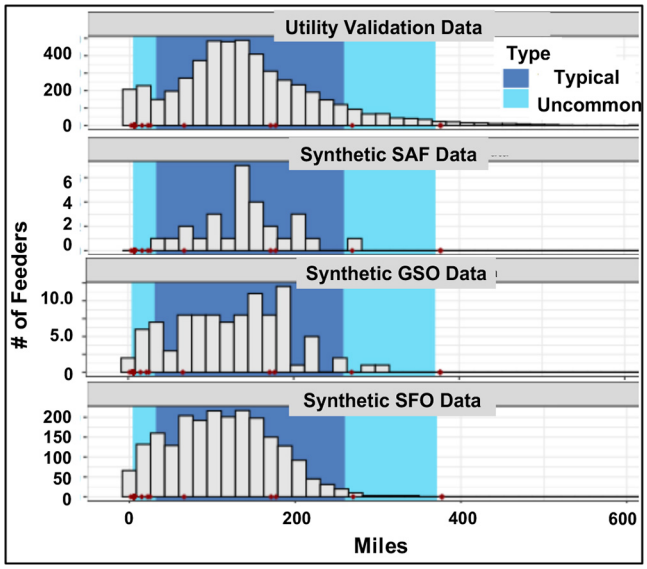


Fig. 15. Graph diameter per feeder (5,020 real utility feeders).

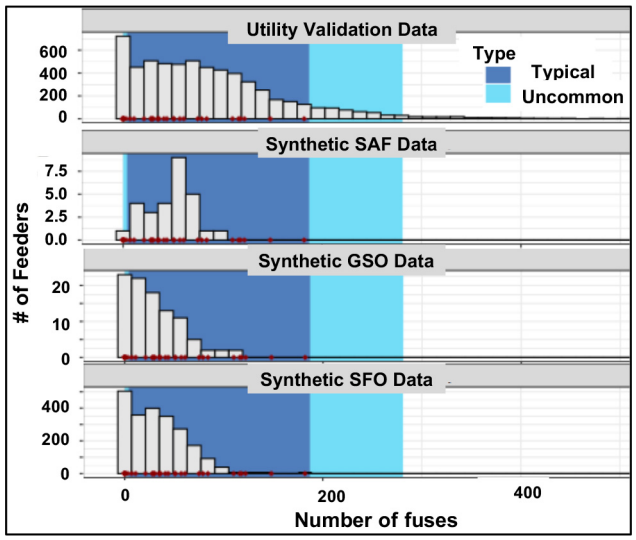


Fig. 16. Number of fuses per feeder (6,013 real utility feeders).

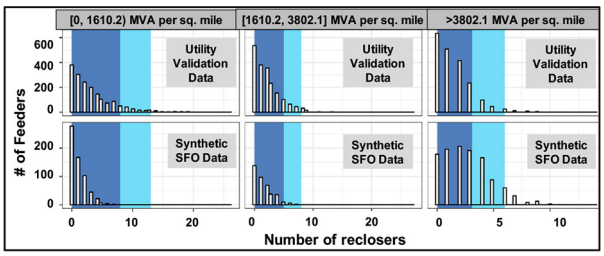


Fig. 17. Number of reclosers per feeder: by load density partitions, with 1735:1930:1990 real utility feeders.

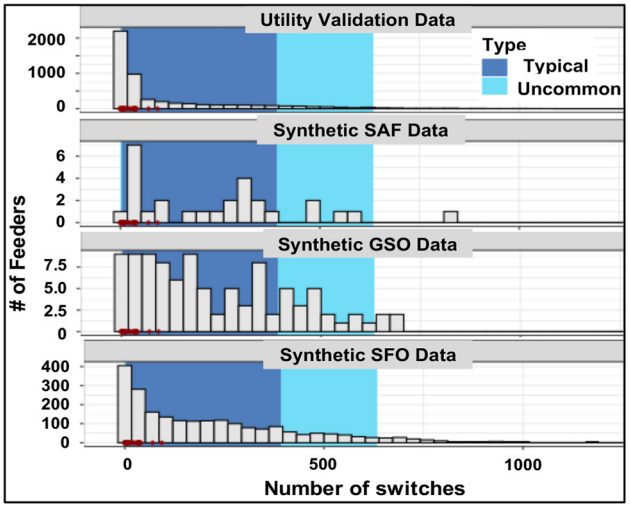


Fig. 18. Number of switches per feeder (5,020 real utility feeders).

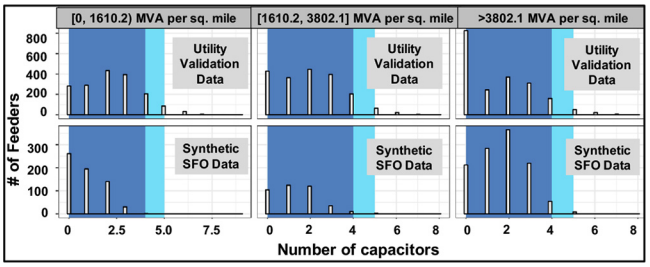


Fig. 19. Number of capacitors per feeder: by load density partitions, with 1735:1930:1990 real utility feeders.

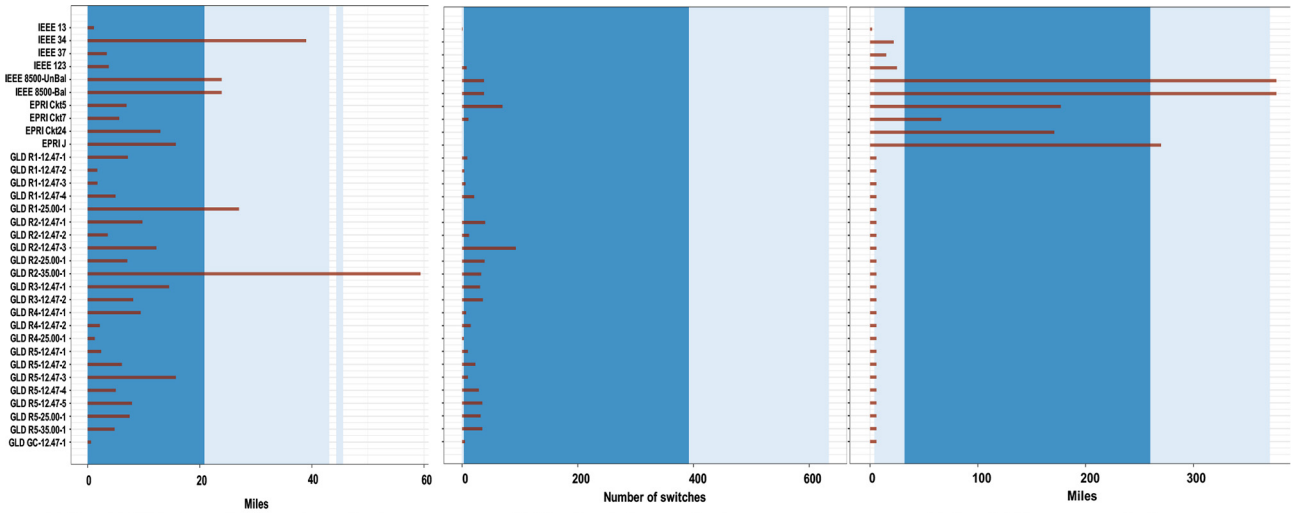
utility distribution. Tables V and VI show a high percentage of single-family units in urban regions connected through several LV lines (as also shown in Fig. 14).

- The validation conclusions are drawn based on the three-pronged strategy that includes expert feedback. Therefore, at times apparent deviations observed between the synthetic and aggregated utility histograms are intentional based on post-processing driven by expert feedback. For instance, Fig. 18 shows a higher number of switches per feeder in the synthetic data. Because SFO has more urban-type feeders, more GOAB (gang-operated air breakers) and manually operated elbow switches were added at appropriate locations during post-processing.
- Finally, Fig. 20 compares selected metrics from the open-access distribution test feeders [4] (IEEE, EPRI, and GridLab-D [GLD] taxonomy feeders) against the validation regions estimated from real utility data (see Table IV). Compared to these test cases with single feeders, the developed large-scale multi-feeder synthetic data sets provide a higher diversity and larger size similar to what is observed in realistic data, thereby promising to push the state of the art in distribution system analysis and optimization algorithms.

In Fig. 16, one needs to note that the synthetic networks have many feeders in urban-type regions and less in rural-type feeders, compared to the aggregated utility data. Therefore, we notice lesser fuses per feeder that are predominantly in shorter single-phase laterals of the urban-type regions.

ACKNOWLEDGMENT

The authors thank the U.S. DOE Advanced Research Projects Agency-Energy (ARPA-E) for their funding support and expert feedback provided through the quarterly review process (Kory Hedman, David Guarrera, and Timothy



(a) Total MV 3-phase line length per feeder (b) Total switches per feeder (c) Graph Diameter per feeder

Fig. 20. Comparison of selected metrics from existing open-access feeders vs. validation region (note, also shown in Figs. 9–19 as red markers in x-axis).

Heidel). We are also grateful for the invaluable support from industry experts, vendors, and data partners: Duke Energy (Leslie Ponder and Lance Fox); Arizona Public Service; City of Loveland (Brieana Reed-Harmel), Colorado; Southern California Edison; Exelon Companies (Steve Steffel; Potomac Electric Power Company (Pepco), Delmarva Power, and Atlantic City Electric); and Electrical Distribution Design (EDD, Jason Bank). Without their support, this work would not have been possible. Additional thanks to our colleagues for valuable inputs: Mike Coddington and Barry Mather (NREL), Claudio Vergara (Off Grid Electric), Roger Dugan (EPRI), Abilash Thakallapelli (General Electric), and Nicolas Gensollen (LiP6, Sorbonne University). The views expressed in the article do not necessarily represent the views of the U.S. Government. The publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

REFERENCES

[1] A. Koirala, L. S.-Ramón, B. Mohamed, P. Arboleya, "Non-synthetic European low voltage test system," *Int. J. Elect. Power Energy Syst.*, vol. 118, Jun. 2020, Art. no. 105712.

[2] *Iowa Distribution Test Systems*. Accessed: Mar. 24, 2020. [Online]. Available: http://wzy.ece.iastate.edu/Testsystem.html

[3] *PGE Prototypical Feeders, Gridlab-D Format*. Accessed: Mar. 24, 2020. [Online]. Available: http://gridlab-d.shoutwiki.com/wiki/PGE_Prototypical_Models

[4] K. P. Schneider *et al.*, "Analytic considerations and design basis for the IEEE distribution test feeders," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3181–3188, May 2018.

[5] F. E. Postigo *et al.*, "A review of power distribution test feeders in the united states and the need for synthetic representative networks," *Energies*, vol. 10, no. 11, p. 1896, 2017.

[6] C. Mateo, T. Gómez, A. Sánchez, J. Peco, and A. Candela, "A reference network model for large-scale distribution planning with automatic street map generation," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 190–197, Feb. 2011.

[7] C. Mateo *et al.*, "Building large-scale U.S. synthetic electric distribution system models," *IEEE Trans. Smart Grid*, to be published, [Online]. Available: https://www.iit.comillas.edu/publicacion/mostrar_publicacion_working_paper.php.es?id=352

[8] F. Postigo *et al.*, "Phase selection algorithms to minimize cost and imbalance in U.S. synthetic distribution systems," *Int. J. Elect. Power Energy Syst.*, to be published, [Online]. Available: https://www.iit.comillas.edu/publicacion/mostrar_publicacion_working_paper.php.es?id=353

[9] C. Coffrin, D. Gordon, and P. Scott. (Aug. 2016). *NESTA: The NICTA Energy System Test Case Archive*. [Online]. Available: https://arxiv.org/pdf/1411.0359.pdf

[10] A. Birchfield *et al.*, "A metric-based validation process to assess the realism of synthetic power grids," *Energies*, vol. 10, no. 8, p. 1233, 2017.

[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] R. Espejo, S. Lumbreras, and A. Ramos, "Analysis of transmission-power-grid topology and scalability, the European case study," *Physica A Stat. Mech. Appl.*, vol. 509, pp. 383–395, Nov. 2018.

[13] R. Espejo, S. Lumbreras, and A. Ramos, "A complex-network approach to the generation of synthetic power transmission networks," *IEEE Syst. J.*, vol. 13, no. 3, pp. 3050–3058, Sep. 2019.

[14] P. Giuseppe, G. Flavia, M. A. Maria, R. P. L. Alexandre, and F. Gianluca, *Distribution System Operators Observatory*, document EUR 27927 EN, Eur. Comm., Brussels, Belgium, 2016, doi: 10.2790/701791.

[15] C. Mateo *et al.*, "European representative electricity distribution networks," *Int. J. Elect. Power Energy Syst.*, vol. 99, pp. 273–280, Jul. 2018.

[16] M. Grzanic, M. G. Flammini, and G. Prettico, "Distribution network model platform: A first case study," *Energies Open Access*, vol. 12, no. 21, p. 4079, 2019.

[17] E. Schweitzer, A. Scaglione, A. Monti, and G. A. Pagani, "Automated generation algorithm for synthetic medium voltage radial distribution systems," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 7, no. 2, pp. 271–284, Jun. 2017.

[18] S. S. Saha, E. Schweitzer, A. Scaglione, and N. G. Johnson, "Framework for generating synthetic distribution feeders using OpenStreetMap," Oct. 2019, [Online]. Available: https://arxiv.org/pdf/1910.07673.pdf.

[19] *National Renewable Energy Laboratory |DR POWER*. Accessed: Mar. 24, 2020. [Online]. Available: https://egriddata.org/search/field_tags/synthetic-data-693/field_topic/distribution-86?sort_by=changed

[20] *BetterGrids.org: A Standards-based Intelligent Repository for Collaborative Grid Model Management*. Accessed: Mar. 24, 2020. [Online]. Available: https://db.bettergrids.org/bettergrids/handle/1001/94

[21] M. Deru *et al.*, "U.S. department of energy commercial reference building models of the national building stock," NREL, Golden, CO, USA, Rep. NREL/TP-550–46481, Feb. 2011.

[22] 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.
[23] "American national standard for electric power systems and equipment—Voltage ratings (60 Hz)," Standard ANSI C84.1-2016, 2011. [Online]. Available: https://www.nema.org/Standards/ComplimentaryDocuments/ANSI%20C84-1-2016%20CONTENTS%20AND%20SCOPE.pdf
[24] R.C. Dugan, M. F. McGranaghan, and H. W. Beaty, *Electrical Power Systems Quality*. New York, NY, USA: McGraw-Hill, 1996.
[25] IEEE Industry Applications Society, *IEEE Recommended Practice for Calculating Short-Circuit Currents in Industrial and Commercial Power Systems*, IEEE Standard 551TM-2006, 2006.
[26] J. J. Burke, *Power Distribution Engineering: Fundamentals and Applications*. New York, NY, USA: Marcel Dekker Inc., 1994.
[27] R. J. Broderick and J. R. Williams, "Clustering methodology for classifying distribution feeders," in *Proc. IEEE 39th Photovolt. Specialists Conf. (PVSC)*, Tampa, FL, USA, 2013, pp. 1706–1710.
[28] J. Cale, B. Palmintier, D. Narang, and K. Carroll, "Clustering distribution feeders in the Arizona Public Service territory," in *Proc. IEEE 40th Photovolt. Specialist Conf.(PVSC)*, Denver, CO, USA, 2014, pp. 2076–2081.
[29] T. A. Short, *Electric Power Distribution Handbook*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2014, p. 26–28.

Venkat Krishnan (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from Iowa State University in 2007 and 2010, respectively. He is a Senior Engineer with the Sensing and Predictive Analytics Group, Power Systems Engineering Center, National Renewable Energy Laboratory. His expertise is in power system operations (markets and variable renewable integration), transmission and distribution grid stability assessment, energy storage integration, long-term capacity expansion planning, and the application of statistical simulations and data analytics methods in power systems.

Bruce Bugbee received the M.S. and Ph.D. degrees in statistics with Colorado State University in 2010 and 2014, respectively. He was a Computational Statistician with National Renewable Energy Laboratory. He is a Principal Data Scientist with Oracle Data Cloud.

Tarek Elgindy received the B.S. degree from the University of Sydney and the M.S. degree in algorithms, combinatorics, and optimization from Carnegie Mellon University. He is an Engineer with the Grid-Connected Energy Systems Modeling Group, National Renewable Energy Laboratory. He has worked extensively on managing large electrical distribution data sets and understanding the impact of network designs on power quality primarily for quasi-static time-series analysis. His research interests include developing market structures for distribution networks, developing synthetic electrical infrastructure data, and understanding the interactions between distribution and transmission systems with high penetrations of distributed energy resources.

Carlos Mateo received the Ph.D. degree in industrial and computer engineering from Comillas Pontifical University, Madrid, Spain, in 2007, where he is a Member of the Institute for Research in Technology, School of Engineering (ICAI). His current research interests include modeling, simulation, and algorithms, especially in the fields of electricity distribution networks, and distributed energy resources.

Pablo Duenas received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Comillas Pontifical University. He is a Research Scientist with the MIT Energy Initiative. In 2014, he joined the Massachusetts Institute of Technology after being a Research Scientist with Comillas Pontifical University. His research focuses on the technical, economic, and regulatory aspects of interdependent natural gas and electricity systems, energy security under carbon-constrained energy policies, and the future of distributed energy resources.

Fernando Postigo received the M.S. degree in industrial engineering with a specialization in electricity from Comillas Pontifical University, Madrid, Spain, in 2016, where he is currently pursuing the Ph.D. degree in power systems with the Institute for Research in Technology. His research focuses on the development of tools for the modeling and analysis of distribution networks in the context of increasing penetrations of renewables.

Jean-Sébastien Lacroix received the B.Sc. and M.Sc. degrees from the Department of Electrical Engineering, Polytechnique Montréal. He was a Power System Engineering Manager with Eaton's CYME International T&D. He has developed power system models and analyses for system distribution planners and operators. His research interests are in power system analysis, modeling, and simulation for distribution systems.

Tomás Gómez San Roman (Senior Member, IEEE) received the Ph.D. degree in industrial engineering from the Polytechnic University of Madrid, Spain, in 1989. He is a Professor of electrical engineering with the Engineering School, Comillas Pontifical University. He has broad industrial experience in joint research projects in the field of electric power systems. From 2011 to 2013, he served as a Commissioner with the Spanish Energy Regulatory Agency. His areas of interest are the operation and planning of transmission and distribution systems, power quality assessment and regulation, and economic and regulatory issues in the electric power sector.

Bryan Palmintier (Senior Member, IEEE) received the B.S. degree in aerospace engineering from Georgia Institute of Technology, the Engineer degree in mechanical engineering and the M.S. degree in aero/astro engineering from Stanford University, and the Ph.D. degree in engineering systems from the Massachusetts Institute of Technology. He is a Manager of the Grid-Connected Energy Systems Modeling Group and a Principal Research Engineer with National Renewable Energy Laboratory. His current research focuses on advanced integrated analysis methods for renewable and distributed energy systems.