IEEE Open Access Journal of
Power and Energy

Received 15 March 2020; revised 19 June 2020 and 21 July 2020; accepted 3 October 2020. Date of publication 7 October 2020;
date of current version 9 November 2020.

Digital Object Identifier 10.1109/OAJPE.2020.3029278

Building Highly Detailed Synthetic Electric Grid Data Sets for Combined Transmission and Distribution Systems

**HANYUE LI[1] (Student Member, IEEE), JESSICA L. WERT[1] (Student Member, IEEE),
ADAM BARLOW BIRCHFIELD[1] (Member, IEEE), THOMAS J. OVERBYE[1] (Fellow, IEEE),
TOMAS GOMEZ SAN ROMAN[2] (Senior Member, IEEE), CARLOS MATEO DOMINGO[2],
FERNANDO EMILIO POSTIGO MARCOS[2], PABLO DUENAS MARTINEZ[3],
TAREK ELGINDY[4] (Member, IEEE), AND BRYAN PALMINTIER[4] (Senior Member, IEEE)**

[1] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA
[2] Institute for Research in Technology, ICAI, Comillas Pontifical University, 28015 Madrid, Spain
[3] Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[4] National Renewable Energy Laboratory, Golden, CO 80401 USA

CORRESPONDING AUTHOR: H. LI (hanyueli@tamu.edu)

This work was supported in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) (Funding provided by the Advanced Research Projects Agency–Energy) under Contract DE-AC36-08GO28308.

ABSTRACT This paper introduces a methodology for building synthetic electric grid data sets that represent fictitious, yet realistic, combined transmission and distribution (T&D) systems. Such data sets have important applications, such as in the study of the wide-area interactions of distributed energy resources, in the validation of advanced control schemes, and in network resilience to severe events. The data sets created here are geographically located on an actual North American footprint, with the end-user load information estimated from land parcel data. The grid created to serve these fictional but realistic loads is built starting with low-voltage and medium-voltage distribution systems in full detail, connected to distribution and transmission substations. Bulk generation is added, and a high-voltage transmission grid is created. This paper explains the overall process and challenges addressed in making the combined case. An example test case, syn-austin-TDgrid-v03, is shown for a 307 236-customer case located in central Texas, with 140 substations, 448 feeders, and electric line data at voltages ranging from 120 V to 230 kV. Such new combined test cases help to promote high quality in the research on large-scale systems, particularly since much actual power system data are subject to data confidentiality. The highly detailed, combined T&D data set can also facilitate the modeling and analysis of coupled infrastructures.

INDEX TERMS Power systems modeling, synthetic power grids, integrated transmission and distribution.

I. INTRODUCTION

THE typical modeling approach for large power systems separates the high-voltage (69 kV and above) transmission grid data set from those of the lower voltage distribution system. Such decoupling at a substation transformer is justifiable for many studies, and it allows transmission system analysis to employ simplified, computationally tractable models for systems where circuit devices can number in the tens of millions. It also enables traditional distribution system analysis to simplify the higher voltage portions of the grid to enable tractable local simulations. But an increasing number of applications show the insights that can be gained by leveraging a massive, data set of a wide-area interconnected transmission grid, combined with the geographic, topological, and electrical configuration for lower voltage distribution substations, feeders, and end users.

Many applications in this paradigm are driven by new developments on medium- and low-voltage networks. Rooftop and larger distributed solar and other distributed generation have multiplied in recent years [1]–[3], and aggregated effects present new opportunities and challenges for both distribution and bulk system planning and operation. Other distributed energy resources (DERs) - such as electric vehicles and controllable load [4], [5], voltage support resources, and distributed storage - are also increasing and offer potential control options for the planning and operation of both the distribution and transmission grid. Further, new monitoring devices, such as advanced metering infrastructure systems and time-synchronized monitors with phasor and waveform analytics [6], can also supply new information for situational awareness at all scales. Resilience studies during severe events also underscore the need to unify transmission grid analysis with a more detailed understanding of the distribution networks they serve. Hurricanes affect both T&D assets, for example, and the preparation and restoration processes can be hindered by a disconnect between T&D data sets [7].

There have also been recent advances in simultaneously modeling T&D systems at scale. For instance, [8] demonstrates a T&D co-simulation framework for market-DER interactions, while [9] presents a modular transmission-distribution-communication framework. Other papers have explored interfacing algorithms including phase-to-sequence coupling [10] and approaches for faster convergence [11].

However, few public test cases exist to enable research in many of these applications, especially those that offer the large-scale, high-fidelity characteristics needed for transmission-distribution interface studies. Ideally, research would be conducted on models of actual grids, but these data sets are not widely available. In the United States, much of these data are considered critical energy infrastructure information, further restricting access. Even when made available to researchers through nondisclosure agreements, much actual grid data cannot be shared publicly [12], [13]. For consumer-level data, data privacy concerns also limit access.

To spur innovation for advancing public research that can be replicated by peers and cross-validated, new synthetic grids have been developed at both the T&D levels. These cases are anchored in a thorough analysis of actual grids, and they start with public, geo-located information for load and generation. In the transmission area, the approach places synthetic substations geographically at the zip code level, assigns voltage levels, and connects them with a transmission line topology matching a combination of electrical, topological, and geographic characteristics [14]–[16]. This approach has been extended for a variety of applications, such as transient stability modeling [17], transmission scenario development [18], and education [19]. In addition to this approach to building synthetic transmission grids, other approaches have been used, such as [20], [21], many of which focus on graph theory and complex network constraints.

At the distribution level, synthetic grids have been constructed by placing secondaries and distribution transformers, and then designing medium-voltage components to support the low-voltage infrastructure [22]. For U.S.-style systems, phase-balancing algorithms are applied on the medium-voltage network to provide a realistically balanced network across three-phase trunks and single-phase laterals [23].

Previous works have also constructed integrated transmission and distribution test cases. An early example was the largely handcrafted system in [24]. More recent efforts have used automated top-down processes. The work of [25] systematically replaced the aggregated load on the transmission-level with duplicated "template" distribution network models and randomized load and DER scenarios to generate the integrated T&D test cases covering the high and medium voltage levels. A top-down method is also proposed in [26], where algorithms are used to build the high voltage, and then selections from existing test systems are added for medium and low voltages. However, all of these approaches result in largely duplicated distribution systems that do not capture the large spatial variations resulting from street and customer locations. They also only represent the 3-phase portion of the networks, which continue down to low voltage in Europe, but are replaced by extensive single-phase laterals in US style medium voltage systems.

Building on the author's previous work of SMART-DS that creates synthetic distribution systems [22], this paper presents a bottom-up methodology to generate data sets that contain a combined synthetic T&D system that is based on actual customer and street locations. Commercially obtained parcel information is utilized to model the load in a highly-detailed manner, where each end-use customer is geographically located, and the load value is estimated according to the size and usage type of each parcel. Next a synthetic distribution network is built from scratch to reflect the common design characteristics for actual distribution systems, while providing an actual design and topology customized for each feeder. Then, on the same geographic footprint, a synthetic transmission system is created to bridge between the generators and the distribution substations.

The bottom-up methodology proposed this paper enables the creation of combined T&D electric grid model that contains highly detailed modeling of electric load and the distribution network, yet covers large-scale geographic footprint. The elements in the combined T&D data set are geographically placed which provides advantages in data realism and enables location-specific studies, e.g. coupled infrastructure studies.

II. DISTRIBUTION SYNTHESIS FOR HIGHLY DETAILED SYNTHETIC GRIDS

The synthetic distribution system is constructed using the U.S. Reference Network Model tool (RNM-US) [22], which adapted the European Reference Network Model tool (RNM) [27] for U.S.-style networks. Fig. 3 shows the main planning stages in the distribution synthesis. RNM-US takes three primary sources of input. The first is a catalog of standard equipment required to build the distribution network (lines, transformers, capacitors, etc.). This includes comprehensive technical parameters (line ampacities, line sections,

FIGURE 1. **Geospatial information for RNM-US.**

transformer kVA ratings, capacitor kvar ratings, etc.) and cost parameters (investment costs and maintenance costs) that are obtained from commercial and open-source data such as [28], [29] or [30]. The second input is OpenStreetMap data, which are used to constrain the layout of power lines and locate service drops, and describe building footprints and heights. The third input is commercially obtained parcel information that describes parcel use categories (single-family or multi-family residential, hotel, hospital, school, industrial, etc.) and load profiles from a database of consumer archetypes for the same categories. Fig. 1 illustrates the result of processing the geospatial information.

A. CONSUMER LOCATION AND DEMAND ESTIMATION

Consumer coordinates are extracted from building information. A parcel category is labeled to each building by spatially intersecting building centroids and parcel polygons. The peak load of each consumer is calculated by assuming that the peak load is correlated with the building volume for the building category. The building volume is estimated by considering each building as a rectangular prism and multiplying the building footprint by its height. A linear interpolation is applied to load data from the database of reference building models (e.g. [31]) to determine the peak customer load as a function of building volume and building-use category. Two additional points are added to the database to facilitate the interpolation: zero volume, zero peak load; and an expert estimation of the peak load of the largest building in the area to establish a saturation (Fig. 2). Although the database contains load profiles, only the peak load is interpolated as this is the value that is employed in the industry to design distribution networks. The peak load is also used to infer the voltage level (high, medium, or low) at which each consumer or building connects.

The calculated peak load is coincidental for all buildings, as each category shares the same reference profile. Following distribution planning practices, a simultaneity factor is applied to each voltage level: 0.4 for low-voltage, and 0.8 for medium-voltage consumers [27]. In addition, a power factor is also applied to each consumer to estimate the reactive power demand: 0.95 for residential and commercial, and 0.98 for industrial consumers. Each consumer is hence represented by its PQ peak load, after applying a simultaneity factor.

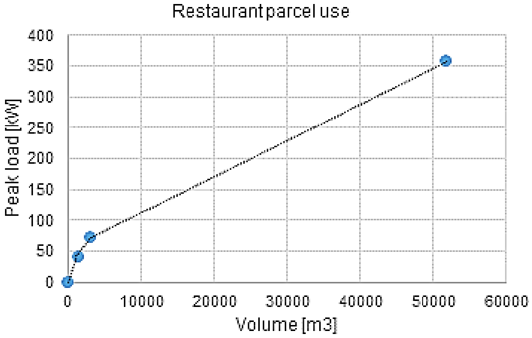

FIGURE 2. **Interpolation example for restaurants.**

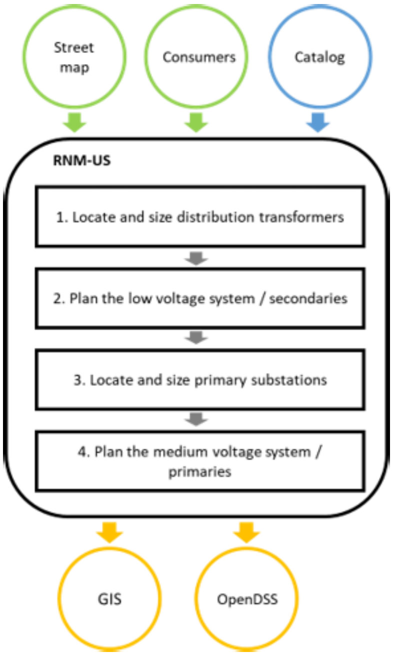

FIGURE 3. **Planning stages in RNM-US.**

Once the peak loads are determined, timeseries load profiles are then attached to each customer. The ResStock [32] and ComStock [33] tools were used to generate thousands of timeseries load profiles that considered a variety of factors such as customer category, building vintage, type of heating used, number of floors, etc. These profiles considered different schedules for cooling/heating systems and accounte for different occupancies of buildings. Each customer then was assigned a Resstock or Comstock profile, which was done by selecting the profile from the same customer category with the maximum load value being closest to the peak determined from the reference profiles. This allowed profiles to have a range of behaviours while still matching the coincident peaks used for planning the size of network equipment.

B. LOCATE AND SIZE DISTRIBUTION TRANSFORMERS

Distribution transformers are located next by identifying clusters of consumers and then locating the transformer for each cluster in nearby streets or right-of-ways. Clusters are built

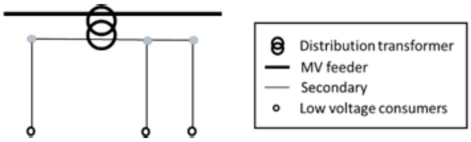

FIGURE 4. **Tree configuration.**

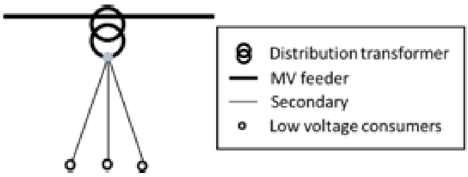

FIGURE 5. **Star configuration.**

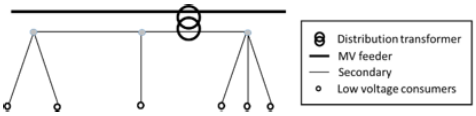

FIGURE 6. **Hybrid configuration.**

starting with the minimum spanning tree that connects all the consumers and then broken into subtrees (clusters) of feasible sizes [34]. The feasibility of a cluster is checked considering thermal and voltage constraints. As the layout is not yet known in this stage, the estimation is made under the assumption that each load is directly connected to the transformer. This allows us to verify that a network solution exists that can supply the consumers with the distribution transformer located in that position. Although this is not the final design, it allows us making a check on feasibility.We use mainly single-phase, center-tap transformers for residential consumers and three-phase transformers for larger loads (e.g., commercial or industrial consumers).

C. PLAN THE LOW-VOLTAGE SYSTEM/SECONDARIES

The secondaries are designed within each cluster to connect the consumers to the respective distribution transformer. Two basic types of configurations are mainly observed in the United States. First, a tree configuration, with the distribution transformer as the root of the tree. A main section branches out with several service drops along the street to connect to consumers (see Fig. 4). Second, a star configuration, where each distribution transformer connects directly to several houses (see Fig. 5). In addition, we use a third hybrid configuration, where the main configuration is a tree, but several star connections might exist, especially for nearby houses (see Fig. 6).

D. LOCATE AND SIZE PRIMARY SUBSTATIONS

The substation planning stage also begins by clustering the loads. But in the case of substations, the loads are the distribution transformers (planned previously) and medium-voltage consumers. The clustering algorithm is the same as that used for distribution transformers: it starts by breaking down the minimum spanning tree and then verifies that the obtained clusters will be feasible in terms of thermal and voltage constraints. In the case of substations, the area covered is much larger than that for distribution transformers,because of the higher nominal voltages and capacities.

E. PLAN THE MEDIUM-VOLTAGE SYSTEM/PRIMARIES

The algorithm for planning medium-voltage feeders again starts by building a minimum spanning tree [35], which represents the shortest solution in the absence of constraints, and then applying a branch-exchange algorithm to obtain a feasible solution that meets thermal and voltage constraints [36]. For sizing power lines, the net present value is computed for each of them, depending on their power flow. The optimal component is such that the net present value is minimized while respecting the thermal limits. This process is explained in [23]. The medium-voltage feeder design considers the three-phase feeder trunks as well as two-phase or single-phase laterals depending on the supplied loads. In addition, the connection of loads to the different phases is made by minimizing imbalance across each feeder [23]. Realistic voltage control is introduced by placing voltage regulators and capacitor banks. The voltage regulators are placed along feeders aiming to respect voltage limits, whereas capacitor banks are sized and located by considering their impact on energy losses. The next step in the medium-voltage system planning stage checks and improves reliability indexes. Switches are installed along feeders by searching for the most critical branches in terms of upstream demand and downstream failure rate. The stopping criterion for switches is based on target metrics, such as the System Average Interruption Duration Index (SAIDI) and the System Average Interruption Frequency Index (SAIFI), which are estimated by simulating how the maintenance crews locate and isolate faults and restore service [31]. Loops with open switches between feeders can also be included to improve reliability until the aforementioned targets are satisfied.

RNM-US outputs the designed distribution system in both OpenDSS and Shapefile formats. This provides a topological geographic information system representation and enables power flow simulations of the planned system. Validation and calibration of these networks was performed by applying a three-pronged approach comprising of (1) statistical validation of characteristic metrics, (2) operation validation of power flow and (3) validation from industry experts [37].

F. POST-PROCESSING

Once RNM-US has created the base network structure, post-processing is performed using the Distribution Transformation Tool (DiTTo) [38] to enhance some details of the base electrical network from RNM-US. These include adding specific control schemes for voltage regulators and capacitors, detailing the multi-transformer bank arrangement inside substations, applying fuse and recloser settings, and attaching time series loads from ResStock$^{\text{TM}}$ and ComStock$^{\text{TM}}$ [32]. Reactive power profiles are estimated based on the time varying breakdown of end uses. Additionally, a rich set of

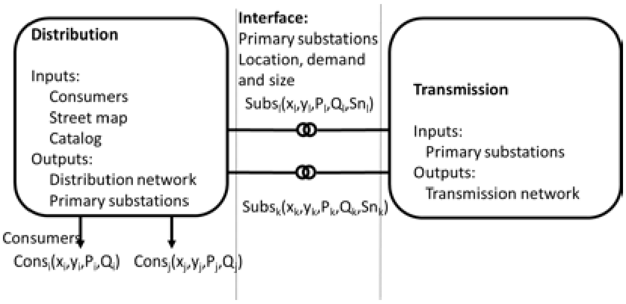

FIGURE 7. **Interface between T&D.**

scenarios is attached to the data sets generated from RNM-US, including various penetrations of solar and storage in OpenDSS as well as several patterns for electric vehicle locations, demand response customers, measurement equipment, controllable switches, and electrical fault locations among others.

The post-processing phase also creates inputs to the transmission grid synthesis phase using a bottom-up approach for the combined synthetic T&D system. Specifically, the distribution system is constructed first, and then post-processing is used to provide aggregated summary inputs at the substation or load bus for the transmission system synthesis. As shown in Fig. 7, the substations are characterized by their location, the demand that they serve—both real and reactive power—and their rated capacity.

III. TRANSMISSION SYNTHESIS FOR HIGHLY DETAILED SYNTHETIC GRIDS

The creation of a synthetic transmission grid consists of three general steps: substation planning, transmission planning, and reactive power planning. Substation planning defines the geographic locations and voltage levels of load and generation nodes in the synthetic transmission system. The load substations here are provided from the distribution system described. Transmission planning connects the nodes with different voltage levels of transmission lines. The transmission line parameters and network topologies are created based on the references of appropriate nominal voltage levels. Reactive power planning sets generator voltage regulations, tap-changing transformers, and shunt capacitors/reactors to provide a realistic distribution of bus voltages and support convergence of the AC power flow solution. The diagram of the synthesis process is shown in Fig. 8.

A. SUBSTATION PLANNING

Substation planning creates nodes in the transmission network. In this step, each substation is initiated with the assignment of geographic coordinates, load, and generation. The substation is then configured internally with buses and transformers according to the assigned nominal voltage levels.

1) LOAD SUBSTATIONS

To ensure a constant T&D data set, the distribution data set synthesized from the previous section is used to create and locate load substations in the synthetic transmission system. In our approach, the subtransmission system (e.g., 69 kV) is included in the transmission synthesis. The load substations in the synthetic transmission system represent an aggregated view of the distribution system. They adopt the geographic coordinates of the distribution substations and summation of coincident loads of end-use customers. As described in section II, the coincident loads are a scaled version of the peak load to account for nonsimultaneous peak loads. In addition, 5% power losses are added when aggregating the distribution loads. The 5% losses value is only used for the interface with the transmission system when aggregating the load of the primary substations for transmission systm design. The value of 5% was obtained by simulating the distribution feeders in OpenDSS and determining the average losses on the distribution system over all feeders. When simulating the complete T&D system, these loss factors are replaced with the actual distribution system and loses are simulated in detail based on electrical properties.

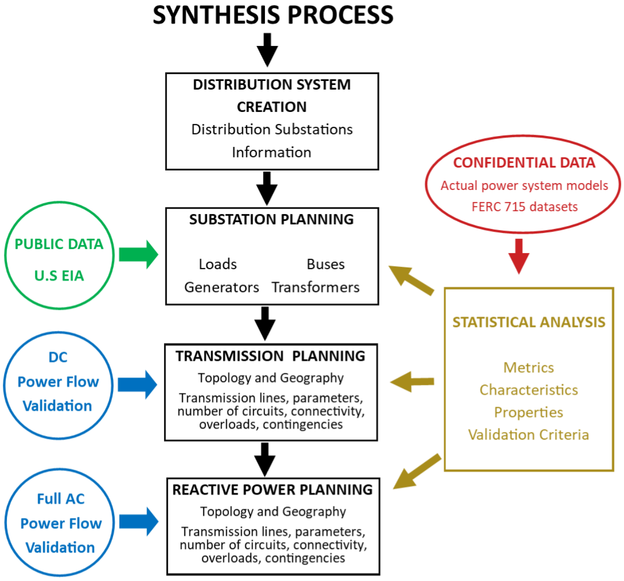

FIGURE 8. **Transmission data set Synthesis Process Diagram.**

2) GENERATION SUBSTATIONS

The U.S. Energy Information Administration conducts an annual survey of the nation's power plants. The 2014 survey data are readily available to the public [39]. These data include the geographic coordinates, fuel type, generation capacity, and number of units of all generation plants. These power plants become the generation for the bulk energy system in the model.

These power plants are then clustered geographically into generation-only substations. Although it is possible to have multiple fuel types within one power plant, a simplifying assumption is made so that hydro, nuclear, and renewable energy resources are not grouped together.

3) SUBSTATION CONFIGURATIONS

The configuration of the substation includes the assignment of voltage levels and the creation of buses and transformers. Given a set of substations sited geographically across a system, the distribution of the voltage level in each area is first determined to reflect the real statistics and system needs.

To assign voltage levels to specific substations, the lowest voltage level is assigned to all substations first, because it typically covers most of the geographic region. Some substations are then upgraded to higher voltage levels using a nonuniform random selection to produce percentages consistent with real networks. Substations with larger load and generation have a higher probability of being assigned with high-voltage levels.

At least one bus is created within a substation for each voltage level assigned. For load substations, one bus is created with the load attached; this bus is the coupling point to the distribution feeders. In generation substations, multiple buses can be created with generators attached, depending on the number of units in each power plant. If multiple voltage levels are assigned to one substation, transformers are also created to step up or step down the voltage.

B. TRANSMISSION PLANNING

Transmission planning connects the buses established in the substation planning stage with the transmission network at multiple nominal voltage levels. The algorithm of creating the synthetic network from [14] is used to generate the transmission lines automatically. In [14], a connected graph is built at each voltage level using Delaunay triangulation, which links buses in a network with transmission lines based on nearest-neighbor concepts. The network is also designed so that the combined graph of all voltage levels can still remain connected even if one substation is removed [14]. This property does not allow for radial substations, and it enhances contingency security. Many real power systems match these properties.

Transmission line electrical parameters required for power flow analysis include series impedance, shunt admittance, and MVA limits. Realistic per-distance parameters are assigned to synthetic lines based on data sheets and references appropriate to the assigned nominal voltage level [14]. Because the network is connected, the created synthetic transmission system at the end of transmission planning stage has a DC power flow solution.

C. REACTIVE POWER PLANNING

The approach from [15] is used in this paper to move incrementally from a DC power flow solution at the end of the transmission planning step, to a full AC power flow solution with a reasonable set of reactive power support devices. This begins by initializing the system to have a very large number of reactive devices controlling the voltage magnitude of most system buses to a common, flat voltage. Then, iteratively, some of the temporary devices are removed at each step, adjusting the remaining ones by repeated AC power flow solutions. Reactive power planning introduces generator voltage regulations, tap-changing transformers, shunt capacitors, and shunt reactors into the synthetic transmission system.

D. WIDE-AREA TRANSMISSION SYSTEM WITH REGIONAL DISTRIBUTION DETAILS

The combined synthetic T&D system can be considered as a self-sustained island. Alternately, the transmission network can be interconnected to a synthetic grid of larger geographic footprint, creating a system in which only some parts of the combined system have distribution details. The load substations in the full synthetic transmission system initially use the geographic coordinates and population of each postal code area obtained from the public U.S. census data set [40], where the load is approximated to be proportional to the population [14]. It is incrementally updated to meet the specific distribution system topology.

IV. EXAMPLE TEST CASE

To demonstrate the methodology developed in this paper, this section presents the syn-austin-TDgrid-v03 test case, its validation, and simulation.

The example case developed using the presented methodology covers the geographic region of Travis County, Texas, including the city of Austin and surrounding areas in central Texas. This data set serves 307,236 customers loads with total system peak of 3,254 MW. There are in total 140 substations in the system, with 69 kV and 230 kV nominal voltage level. This data set includes a mix of 448 rural, suburban, and urban feeders, and 132,406 distributed transformers. There are in average 5.3 consumers per distribution transformer; the distribution transformer capacities are in the range 10-1500 kVA, and ANSI ratings are used for the maximum allowed voltage range.

The synthetic distribution network built models diversity in the following terms: 1) There are urban, suburban, and rural circuits in the data sets adapted to the different characteristics and dispersion of consumers. In particular, the urban/suburban and rural circuits have different design targets, for example, network length and reliability. 2) Several distribution nominal voltage levels are considered, specifically 4kV, 12.47kV, and 25kV. 3) Several approaches for voltage management are considered: voltage regulators and/or capacitor banks. 4) The loading of the network components depends on the discrete network components available in the input catalog.

Summary overview statistics of the case are given in Table 1. It is important to note that while the load is realistically modeled, the electric network that supplies the load in this synthetic test case is intentionally designed to be different from the actual system on the same geographic footprint. This prevents the synthetic data set from revealing critical energy infrastructure information, but still provides the users realistic test cases to develop techniques that can be applied to the real system. This test case is publicly available for download at [41].

TABLE 1. Overview statistics of the syn-austin-TDgrid-v03 test case.

Customer loads	307,236
Total peak load	3254 MW
69 kV substations	119
230 kV substations	21
Generator units	39
Feeders	448
Feeder length	19,726 km
Distributed transformers	132,406
69 kV transmission lines	229
230 kV transmission lines	34
Transmission buses	160
Distribution electric nodes	1,654,691

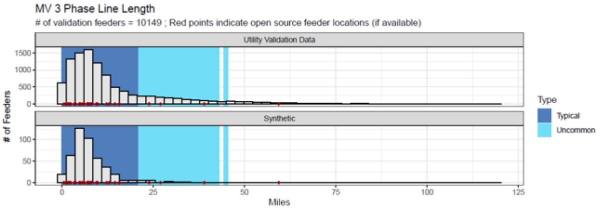

FIGURE 9. Comparison of total feeder line lengths for real utility networks and synthetic Austin data set.

The combined T&D networks are presented in Fig. 12. Here, the blue lines are the 230 kV transmission network, and the green lines are the 69 kV transmission network. For the distribution grid, the pink, red, orange, and yellow lines show distribution feeders. Within the footprint of this case, urban and residential areas are reflected in the system topology. Fig. 13 offers a closer look at the downtown area of the synthetic test case.

A. VALIDATION

The base synthetic system was then validated using previously established techniques developed by the authors. Specifically, the distribution system was statistically validated as described in [37]. This includes comparisons of several key structural properties of the network that were compared to comprehensive data obtained from tens of thousands of actual feeders from several U.S. utilities and found to be within statistically comparable ranges. Visual representations of the differences for two selected metrics are illustrated in Fig. 9 and 10. Operational validation was also performed by simulating powerflow using OpenDSS and demonstrating that voltages remain within ANSI standards for the base (no DER) case. We also compare voltage profiles and overall distributions of voltages to real system data, e.g. 11.

The transmission system was validated as described in [42], which presents validation metrics based on features (e.g. system topology, parameter values) observed in the Eastern Interconnection, the Western Interconnection, and smaller subset cases from each. This process is based on geography and accounts for a wide range of design practices in actual grids. The system metrics are presented in Table 2.

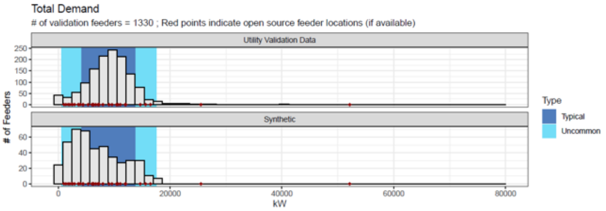

FIGURE 10. Comparison of total feeder loads for real utility networks and syn-austin-TDgrid-v03.

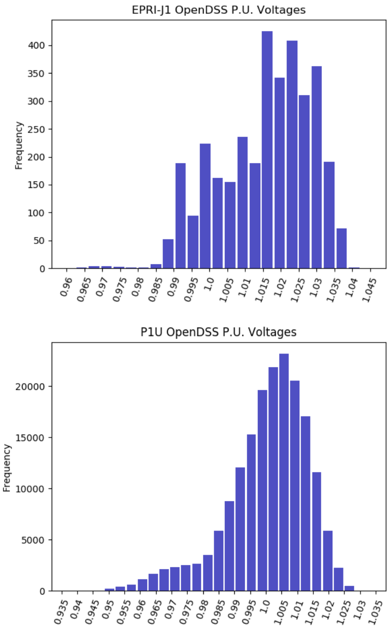

FIGURE 11. Comparison of voltage histograms from public real feeder (EPRI J1, top), vs. syn-austin-TDgrid-v03 (bottom). Our synthetic data is smoother since it contains hundreds of feeders.

For each system metric, the synthetic system's values are compared to the ranges found in the studies performed in [42]. The values match the criteria well with the the exception of slight deviation from four validation metrics. These deviations can be attributed to higher level modeling representations of buses within substations than what might be present in the built grid and the small size of the syn-austin-TDgrid-v03 test case transmission network. This validation provides support that the synthetic network built from a bottom-up approach serves as a realistic representation of an electric grid.

B. COMBINED T&D SIMULATION

To provide a truly combined T&D test system, it is also necessary to be able to simulate the combined power flow and conduct other analyses; however, existing individual tools are not well suited for this combination or large scale. Some of the challenges include: Transmission power flow and dynamics tools typically capture positive sequence, whereas distribution tools use full three-phase, unbalanced models.

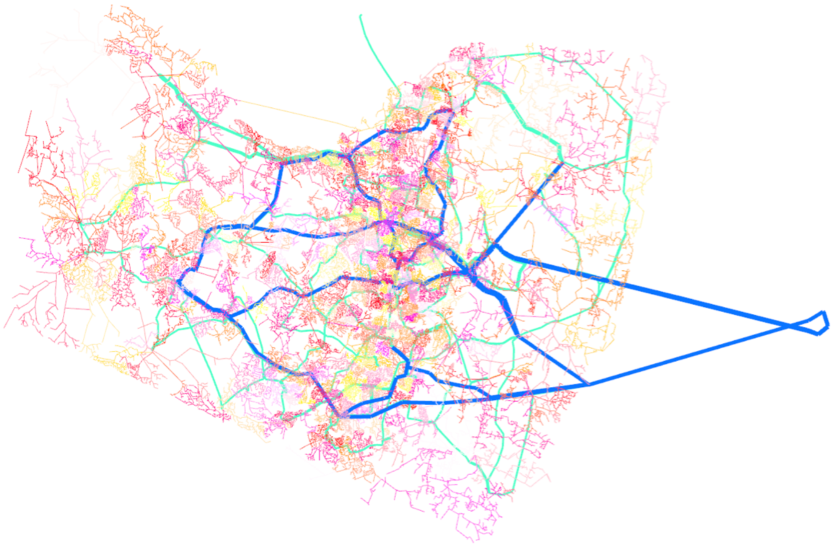

FIGURE 12. The syn-austin-TDgrid-v03 test case, geo-located in Travis County, Texas. Blue and green lines are the transmission grid, with other lines showing distribution feeders. This is a synthetic test case that does not represent any actual grid.

TABLE 2. Validation of the transmission-level network.

#	Validation Metric	Criteria	Austin Case
1	Buses per substation	Mean 1.7-3.5 Exponential Decay	1.24
2	Percent of substations containing buses in kV Range	≤200 kV, 85-100% >201 kV, 7-25%	100% 15%
3	Substations with load	75-90%	91.40%
4	Load per bus	Mean 6-18 MW Exponential Decay	18.8 MW
5	$\frac{GenerationCapacity}{Load}$	1.2-1.6	1.17
6	Substations with Generators	5-25%	8.57%
7	Generator Capacities	25-200 MW, 40+% 200+ MW, 5-20%	64.10% 15.38%
8	Committed Generators	60-80%	76.92%
9	Generators Dispatched above 80% Capacity	50+%	90.00%
10	Generator $\frac{MaxQ}{MaxP}$	0.40-0.50, >70% Voltage Level Number of buses at that level	100.00% 230 21
11	Transformer per-unit X, own base	80% within [0.05, 0.2]	100.00%
12	Transformer X/R ratio and MVA limits, by kV level	X/R 40% below median MVA 40% below median X/R 40% above median MVA 40% above median X/R 80% within 10-90 range MVA 80% within 10-90 range	46.70% 30.00% 53.30% 70.00% 100.00% 100.00%
13	Line per-unit, per-distance reactance, by kV level	70% within 10-90 range	100.00%
14	Line X/R ratio and MVA limits, by kV level	X/R 70% within 10-90 range MVA 70% within 10-90 range	96.60% 100.00%
15	$\frac{Lines}{Substations}$, by kV level	1.1-1.4	1.38
16	Lines on min. spanning tree	45-55%	55.20%
17	Distance of line along Delauney triangulation, by kV level	1, 65-85% 2, 15-25% 3+, 3-10%	86.20% 10.30% 3.40%
18	Total line length / MST	1.2-2.2	2.1

Few existing tools are well suited for solving >1 million electrical nodes at once, or for tackling even larger T&D test systems now under development. And off-the-shelf tools don't simultaneously support bulk generator models, controls, and dispatch while also capturing the specifics of distribution-specific voltage control devices and controls.

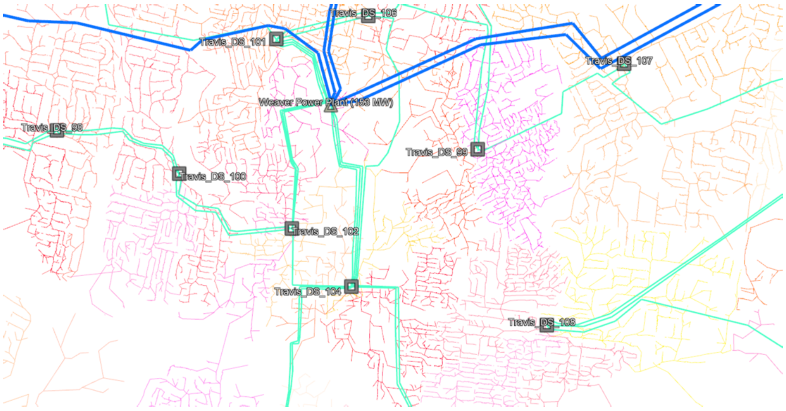

FIGURE 13. **Zoomed-in view of downtown area in the syn-austin-TDgrid-v03 test case, geo-located in Travis County, Texas. Blue and green lines are the transmission grid, with other lines showing distribution feeders. This is a synthetic case that does not represent an actual grid.**

To overcome this challenge, the Austin data set includes instructions and files to be able to run in multiple ways: transmission-only using Powerworld, distribution-only using OpenDSS, and as a co-simulation that brings together Powerworld and OpenDSS to create a combined T&D simulation. The co-simulation uses the open-source HELICS$^{\text{TM}}$ co-simulation [9] framework that also makes it easy to add in other tools for extended analyses, such as the inclusion of an advanced control/optimization scheme, or rich model of other infrastructures, such as transportation.

V. CONCLUSION AND FUTURE WORK

This paper presents pioneering work to create a synthetic combined transmission-distribution power system data set that covers large geographic regions while providing highly detailed models down to the electric meter-level. It builds on past methods developed for separate synthetic distribution and transmission data sets. The distribution system is constructed using the RNM-US, which takes input from OpenStreetMaps data, parcel information, as well as a catalog of standard distribution network equipment. On the same geographic footprint as the distribution system, a synthetic transmission system is developed using a bottom-up approach, where the interface between T&D is the set of substations determined by the distribution system. Using Delaunay triangulation, a synthetic transmission network is created to connect load and generator substations with different voltage levels of transmission lines. Reactive power devices are also included in the synthetic data set to achieve the realism of bus voltage distribution and the convergence of the AC power flow solution. Because the input data of the synthesis process is not confidential, the combined T&D data set is publicly available, and it can be shared freely.

This approach and resulting data set, represent the first large-scale, publicly available T&D data set. It provides an advanced and realistic testing platform to promote next-generation research on large-scale systems, particularly considering the rapid growth in DERs. The combined data set also enables the potential of coupled infrastructure studies. Because the synthetic distribution system model extends to end-use customers, geographic coordinates of electric meters and distribution feeders are available at a high resolution and can facilitate coupling between the power system and other infrastructures, such as the transportation network. This enables studies on the impact of high penetrations of electric vehicles on the power systems and the impact of dramatic grid changes (e.g., a blackout) on transportation systems using coupled models of electric and transportation infrastructures.

Ongoing future work aims to apply these techniques to build even larger T&D data sets consistent with larger utility and/or independent system operators footprints to enable full simulation of T&D at scale. The future work also plans to create scenarios representing a wide spectrum of load levels, DER penetration levels, and bulk generation fuel mix for the combined synthetic T&D system. Those scenarios can enable the research capability of determining controls and hardware upgrades for effective T&D operations.

ACKNOWLEDGMENT

A portion of the research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and 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] M. Begovic, A. Pregelj, A. Rohatgi, and D. Novosel, ''Impact of renewable distributed generation on power systems,'' in *Proc. 34th Annu. Hawaii Int. Conf. Syst. Sci.*, Jan. 2001, pp. 654–663.

[2] T. Adefarati and R. C. Bansal, "Integration of renewable distributed generators into the distribution system: A review," *IET Renew. Power Gener.*, vol. 10, no. 7, pp. 873–884, Aug. 2016.

[3] B. Palmintier *et al.*, "On the path to SunShot: Emerging issues and challenges in integrating solar with the distribution system," Nat. Renew. Energy Lab., NREL, Golden, CO, USA, Tech. Rep. NREL/TP-5D00-65331, May 2016. [Online]. Available: http://www.nrel.gov/docs/fy16osti/65331.pdf

[4] G. A. Putrus, P. Suwanapingkarl, D. Johnston, E. C. Bentley, and M. Narayana, "Impact of electric vehicles on power distribution networks," in *Proc. IEEE Vehicle Power Propuls. Conf.*, Sep. 2009, pp. 827–831.

[5] T. Broeer, J. Fuller, F. Tuffner, D. Chassin, and N. Djilali, "Modeling framework and validation of a smart grid and demand response system for wind power integration," *Appl. Energy*, vol. 113, pp. 199–207, Jan. 2014.

[6] J. A. Wischkaemper, C. L. Benner, B. D. Russell, and K. Manivannan, "Application of waveform analytics for improved situational awareness of electric distribution feeders," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 2041–2049, Jul. 2015.

[7] M. Panteli, C. Pickering, S. Wilkinson, R. Dawson, and P. Mancarella, "Power system resilience to extreme weather: Fragility modeling, probabilistic impact assessment, and adaptation measures," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3747–3757, Sep. 2017.

[8] B. Palmintier *et al.*, "IGMS: An integrated ISO-to-appliance scale grid modeling system," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1525–1534, May 2017.

[9] B. Palmintier, D. Krishnamurthy, P. Top, S. Smith, J. Daily, and J. Fuller, "Design of the HELICS high-performance transmission-distribution-communication-market co-simulation framework," in *Proc. Workshop Modeling Simulation Cyber-Phys. Energy Syst. (MSCPES)*, Apr. 2017, pp. 1–6.

[10] Q. Huang and V. Vittal, "Integrated transmission and distribution system power flow and dynamic simulation using mixed three-sequence/three-phase modeling," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3704–3714, Sep. 2017.

[11] G. Krishnamoorthy and A. Dubey, "Transmission–distribution cosimulation: Analytical methods for iterative coupling," *IEEE Syst. J.*, vol. 14, no. 2, pp. 2633–2642, Jun. 2020.

[12] (2003). *Federal Energy Regulatory Commission (FERC). FERC Order 630-A: Critical Energy Infrastructure Information*. [Online]. Available: https://www.ferc.gov/legal/maj-ord-reg/land-docs/ceii-rule.asp

[13] *Enhancing the Resilience of the Nation's Electricity System*. National Academies Press, Washington, DC, USA, 2017.

[14] 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.

[15] 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.

[16] A. B. Birchfield, H. Li, and T. J. Overbye, "Security considerations in transmission planning for creating large synthetic power grids," in *Proc. Clemson Univ. Power Syst. Conf. (PSC)*, Sep. 2018, pp. 1–4.

[17] T. Xu, A. B. Birchfield, and T. J. Overbye, "Modeling, tuning, and validating system dynamics in synthetic electric grids," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6501–6509, Nov. 2018.

[18] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," in *Proc. IEEE Texas Power Energy Conf. (TPEC)*, Feb. 2018, pp. 1–6.

[19] A. B. Birchfield, T. J. Overbye, and K. R. Davis, "Educational applications of large synthetic power grids," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 765–772, Jan. 2019.

[20] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating statistically correct random topologies for testing smart grid communication and control networks," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 28–39, Jun. 2010.

[21] S. Soltan, A. Loh, and G. Zussman, "A learning-based method for generating synthetic power grids," *IEEE Syst. J.*, vol. 13, no. 1, pp. 625–634, Mar. 2019.

[22] C. Mateo *et al.*, "Building large-scale U.S. Synthetic electric distribution system models," *IEEE Trans. Smart Grid*, early access, Jun. 19, 2020, doi: 10.1109/TSG.2020.3001495.

[23] F. Postigo *et al.*, "Phase-selection algorithms to minimize cost and imbalance in U.S. synthetic distribution systems," *Int. J. Electr. Power Energy Syst.*, vol. 120, Sep. 2020, Art. no. 106042.

[24] R. Billinton and S. Jonnavithula, "A test system for teaching overall power system reliability assessment," *IEEE Trans. Power Syst.*, vol. 11, no. 4, pp. 1670–1676, Nov. 1996.

[25] N. Pilatte, P. Aristidou, and G. Hug, "TDNetGen: An open-source, parametrizable, large-scale, transmission, and distribution test system," *IEEE Syst. J.*, vol. 13, no. 1, pp. 729–737, Mar. 2019.

[26] M. Sarstedt, S. Garske, C. Blaufus, and L. Hofmann, "Modelling of integrated transmission and distribution grids based on synthetic distribution grid models," in *Proc. IEEE Milan PowerTech*, Jun. 2019, pp. 1–6.

[27] C. M. Domingo, T. G. S. Roman, A. Sanchez-Miralles, J. P. P. Gonzalez, and A. C. Martinez, "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.

[28] *Per-Unit Cost Guide*. Accessed: Mar. 2020. [Online]. Available: http://www.caiso.com/informed/Pages/StakeholderProcesses/ParticipatingTransmissionOwnerPerUnitCosts.aspx

[29] *Bare Overhead Conductors*. Accessed: Mar. 2020. [Online]. Available: https://www.nexans.us/eservice/US-en_US/fileLibrary/Download_540190082/US/files/BareOverhead

[30] *IEEE Application Guide for AC High-Voltage Circuit Breakers Rated on a Symmetrical Current Basis*, Standard IEEE Std. C37.010, 1999.

[31] M. Deru *et al.*, "U.S. Department of energy commercial reference building models of the national building stock," National Renewable Energy Lab. (NREL), Golden, CO, USA, Tech. Rep. NREL/TP-5500-46861, Feb. 2011. [Online]. Available: https://www.osti.gov/biblio/1009264

[32] E. J. Wilson, C. B. Christensen, S. G. Horowitz, J. J. Robertson, and J. B. Maguire, "Energy efficiency potential in the U.S. Single-family housing stock," Nat. Renew. Energy Lab. (NREL), Golden, CO, USA, Tech. Rep. NREL/TP-5500-68670, Dec. 2017. [Online]. Available: https://www.osti.gov/biblio/1414819/

[33] E. K. O. Present, "End use load profiles for the U.S. Building stock," National Renew. Energy Lab. (NREL), Golden, CO, USA, Tech. Rep. NREL/PR-5500-75282, Oct. 2019. [Online]. Available: https://www.osti.gov/biblio/1572770

[34] P. Wen and L. Wenxia, "Niche genetic algorithm and minimum spanning tree for substation planning," in *Proc. WRI Global Congr. Intell. Syst.*, 2009, pp. 61–65.

[35] S. N. Ravadanegh, A. Vahidnia, and H. Hatami, "On optimal design and expansion of electrical power distribution systems," *J. Circuits, Syst. Comput.*, vol. 19, no. 1, pp. 45–58, Feb. 2010.

[36] J. Salehi and M.-R. Haghifam, "Long term distribution network planning considering urbanity uncertainties," *Int. J. Electr. Power Energy Syst.*, vol. 42, no. 1, pp. 321–333, Nov. 2012.

[37] V. Krishnan *et al.*, "Validation of synthetic U.S. electric power distribution system data sets," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4477–4489, Sep. 2020.

[38] *Ditto: The Distribution Transformation Tool*. Accessed: Mar. 2020. [Online]. Available: https://github.com/NREL/ditto

[39] (2014). *U.S. Energy Information Administration, Form EIA-860*. Accessed: Mar. 2020. [Online]. Available: http://www.eia.gov/electricity/data/eia860/index.html

[40] U.S. Census Bureau. *2010 Census Gazetteer Files: ZIP Code Tabulation Areas*. Accessed: Mar. 2020. [Online]. Available: https://www.census.gov/geo/mapsdata/data/gazetteer2010.html

[41] *Electric Grid Test Case Repository—Combined T+D Synthetic Dataset*. Accessed: Mar. 2020. [Online]. Available: https://electricgrids.engr.tamu.edu/combined-td-synthetic-dataset/

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

HANYUE LI (Student Member, IEEE) received the B.Sc. degree in electrical engineering from the Illinois Institute of Technology, Chicago, IL, USA, in 2016, and the M.Sc. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2017. She is currently pursuing the Ph.D. degree in electrical engineering with Texas A&M University, College Station, TX, USA.

JESSICA L. WERT (Student Member, IEEE) received the B.S. degree in engineering sciences from Smith College, Northampton, MA, USA, in 2018. She is currently pursuing the Ph.D. degree in electrical engineering with Texas A&M University, College Station, TX, USA.

ADAM BARLOW BIRCHFIELD (Member, IEEE) received the B.E.E. degree from Auburn University, Auburn, AL, USA, in 2014, the M.S. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2016, and the Ph.D. degree in electrical engineering from Texas A&M University (TAMU), College Station, TX, USA, in 2018.

THOMAS J. OVERBYE (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the University of Wisconsin–Madison, Madison, WI, USA. He is currently a TEES Distinguished Research Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.

TOMAS GOMEZ SAN ROMAN (Senior Member, IEEE) is currently a Full Professor of electrical engineering with the Engineering School (ICAI), Universidad Pontificia Comillas, Madrid. His research interests include operation and planning of transmission and distribution systems, power quality assessment, and economics and regulation of the power sector.

CARLOS MATEO DOMINGO is currently a Member with the School of Engineering (ICAI), Institute for Research in Technology (IIT), Universidad Pontificia Comillas, Madrid. His research interests include distribution network models and distributed energy resources.

FERNANDO EMILIO POSTIGO MARCOS is currently pursuing the Ph.D. degree in power systems with the Institute for Research in Technology (IIT), Comillas Pontifical University, Madrid. His research interests include development of tools for the modeling and analysis of distribution networks in a context of increasing penetration of renewables.

PABLO DUENAS MARTINEZ received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Comillas Pontifical University. He joined the Massachusetts Institute of Technology in 2014 after being a Research Scientist at Comillas Pontifical University. He is currently a Research Scientist with the MIT Energy Initiative. His research interests include 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.

TAREK ELGINDY (Member, IEEE) received the B.Sc. degree from The University of Sydney and the master's degree in algorithms, combinatorics, and optimization from Carnegie Mellon University. His research interests include developing markets structures for distribution systems, developing tools for managing and cleaning electrical infrastructure data, and understanding the interaction between distribution and transmission systems with high penetrations of DERs.

BRYAN PALMINTIER (Senior Member, IEEE) received the B.S. degree in aerospace engineering from the Georgia Institute of Technology, the M.S. degree in aero/astro engineering and the degree in mechanical engineering from Stanford University, and the Ph.D. degree in engineering systems from the Massachusetts Institute of Technology. He is currently the Manager of the Grid-Connected Energy Systems Modeling Group and a Principal Research Engineer with the National Renewable Energy Laboratory.