Experiences developing large-scale synthetic U.S.-style distribution test systems

Bryan Palmintier, Tarek Elgindy
Power systems Engineering Center
NREL
Golden, Colorado, U.S.A.

Carlos Mateo, Fernando Postigo,
Tomás Gómez, Fernando de Cuadra
Universidad Pontificia Comillas
Madrid, Spain

Pablo Duenas Martinez
MIT Energy Initiative
MIT
Cambridge, Massachusetts, U.S.A.

***Abstract*—This paper describes computational, data management, and other experiences developing large-scale, realistic-but-not-real U.S.-style distribution test systems for the Smart-DS project. These test systems cover entire metropolitan areas and include everything from low-voltage secondaries to sub-transmission for hundreds or thousands of feeders making them as much as three orders of magnitude larger than existing single feeder test systems. Lessons learned with automation and data handling are shared to aid data set users and synthetic test grid creators.**

***Index Terms*—Power distribution; power system planning; Reference Network Model; synthetic networks; test systems.**

I. Introduction

Power grid modernization and wide-spread distributed energy resource (DER) integration are driving an explosion of interest in advanced algorithms. However, there has historically been a lack of open-access realistically sized and detailed datasets—particularly for the distribution system. [1]

A number of recent efforts have created large-scale open-source transmission test systems. For instance, an IEEE task force recently released a curated collection of test systems for testing transmission AC optimal power flow (ACOPF) algorithms [2] while other test systems have included data for use in geomagnetic disturbance simulations [3]. Efforts have also been made to refine the creation process for synthetic transmission datasets including load data development [4], network creation algorithms [5], and validation [6].

Past distribution test systems, e.g. [7], [8], have largely focused on only a single medium voltage (MV) feeder, appropriate for simulation engine testing and perhaps approximate system-wide estimates. However, existing data has neither the scale nor complete descriptions needed to simulate entire real-world distribution utility service areas or evaluate emerging, system-wide algorithms, such as architectures and controls for DER support for both distribution and transmission grid services.

To overcome these challenges, the Synthetic Models for Advanced, Realistic Testing: Distribution systems and Scenarios (Smart-DS) project [9] has developed multiple large-scale realistic-but-not-real data sets for complete, synthetic distribution systems for the Santa Fe, New Mexico (NM); Greensboro, North Carolina (NC); and San Francisco, California (CA) metro areas. These datasets are based on actual building and street locations, but with the electrical networks synthesized from the ground up, as if a different collection of utilities were to have re-built the system. They include complete electrical models—customers, LV secondaries, transformers, MV lines, substations, HV sub-transmission, utility equipment, control settings, etc.—plus multiple detailed scenarios—annual time series of load, weather, and solar data; multiple penetrations of solar; locations of other DERs; etc.

This paper provides a brief summary and then takes a close look at the experiences and lessons learned creating these datasets. Specifically, Section II provides a brief overview of the steps used to create the datasets, Section III provides a summary of the datasets including example power flow results, while Section IV provides the core contribution of this paper by describing experiences and lessons learned with an eye toward accelerating future efforts in this area by describing how various key challenges were overcome. Section V concludes with a forward-looking discussion.

II. Creating Large Distribution Test Systems

As seen in Figure 1, our approach for building synthetic distribution datasets follows a sequence of steps. They are summarized in this section below for completeness. In-depth descriptions can be found in [10].

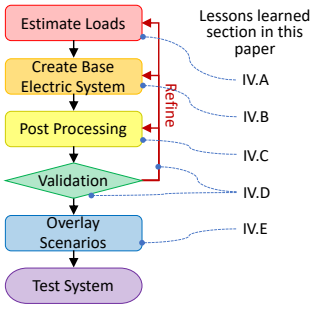

Figure 1: Steps for Creating Smart-DS distribution test systems with a reference to lessons learned and experiences sections in this paper.

A. *Estimating Loads*

The creation process starts by obtaining customer-level information about building location, footprint, and height, along with land or parcel use data for the geographic region of interest. Each building represents a consumption point and the corresponding peak load is estimated in three steps: 1) the building use (e.g., single- or multi-family house, hotel, hospital, industry, school, restaurant, etc.) is determined from the parcel data; and 2) the peak load of each consumer type is obtained using a database of building reference models (e.g.[11]) and then linearly interpolated assuming that the energy consumption is proportional to building volume; and 3) the individual customer loads are scaled down based on a simultaneity factor to produce the planning real and reactive power demands. Both the location and planning load for each consumer are inputs to the Reference Network Model (RNM).

B. *Base electrical system creation*

The base electrical model is then built using RNM, a tool for planning distribution networks, which is able to automatically generate electric distribution network designs for connecting a given set of consumers, based on their location and demand, as well as with the corresponding layout of the streets.[12] This tool can be applied to plan very large-scale distribution areas, covering tens of millions of electrical nodes, and millions of consumers. The networks designed by the model cover three voltage levels, including low voltage secondaries, medium voltage, and sub-transmission.

For this project, a new RNM-US version was developed, that captures the fact the U.S network have quite different designs compared to European networks, notably the use of single (or split) phase 120/240V service to smaller customers, much shorter low voltage connections (secondaries) serving only a few customers, correspondingly smaller distribution transformers, extensive use of single-phase "laterals" at the medium voltage level, and the use of auto transformer voltage "regulators" on many longer medium voltage lines.

C. *Electric System Postprocessing*

As described further in Section IV.B.4), the base RNM-US network, while electrically and topologically complete, is missing some key technical aspects like control settings and other elements. These and other refinements are made during the post-processing stage.

D. *Validation and Refinement*

As detailed in [13], the models then undergo a three-part validation that includes statistical comparisons to real-world utility system data, input from utility and vendor experts, and operational validation—where power flow results are reviewed. This process is conducted iteratively with the results of the various validation components used to identify bugs, inform modifications to the core RNM-US algorithms, input catalogs, and/or postprocessing steps to ensure the synthetic distribution systems provide a realistic representation consistent with those found in real distribution systems. Operational metric bounds were used to identify changes in RNM or post-processing changes such as updating the setpoints for voltage regulating equipment.

E. *Scenario overlay development*

As a final step, detailed timeseries scenarios are generated and attached to the dataset. For load, the single period "planning" load model is augmented with year-long, 15-minute timeseries load profiles for every customer made by sampling from thousands of region appropriate residential and commercial building prototypes bottom-up simulation in EnergyPlus/Open Studio included in ResStock [14] and ComStock [15]. Reactive power is estimated based on typical end-use power factors and the resulting loads are attached to the buildings based on building category and peak yearly load. In addition, richly detailed scenarios for distributed energy resources (DERs) and other uses cases are also added.

III. SUMMARY OF SMART-DS DATASETS

Using this procedure, we have built three synthetic test systems [9]. These cover metropolitan areas in the U. S. and are named using the corresponding 3-letter airport abbreviation:

- **Santa Fe, NM (SAF)**, the smallest dataset. It covers the urban/sub-urban area and is still quite large for distribution test systems: it covers eight substations and twenty-eight feeders; (Figure 2);

Figure 2: Footprint of the synthetic Santa Fe, NM (SAF) distribution system

- **Greensboro, NC (GSO)**, the mid-sized dataset that is about 2.5x larger than SAF and explicitly separates the metro area into three regions: urban/suburban, rural, and industrial-heavy (Figure 3); and

Figure 3: Footprint of the synthetic Greensboro, NC (GSO) distribution system

- **San Francisco Bay Area, CA (SFO)**, the largest dataset at roughly 30x larger than GSO and covering multiple cities and surrounding suburban and rural areas. The SFO dataset explicitly includes a variety of system designs such as older 4kV regions, 25kV rural areas, both wye and delta configured phasing, and various voltage control approaches with and without regulators (Figure 4).

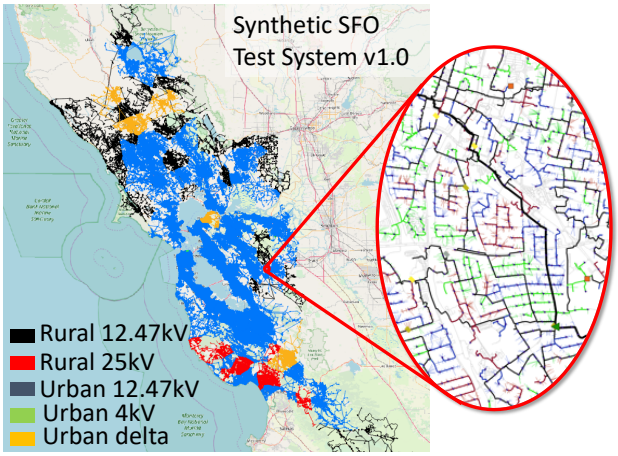

Figure 4: The synthetic greater San Francisco Bay Area (SFO) distribution system [16]

A summary of size and other metrics for each of these datasets is included in Table I. Note that in this table, a "bus" refers to any electrical junction and may have one to three phases within the same bus. In contrast, "nodes" counts each phase at a junction separately.

The statistical portion of our 3-part validation is summarized in Table II. For each listed metric, it lists the percentage of feeders for each test system that fall in the typical (Typ), uncommon (UnC), and rare ranges derived from our validation dataset of thousands of actual feeder models

TABLE I: SUMMARY OF THE SMART-DS DATASETS (V1.0)

	SAF	GSO	SFO
Buildings	38,589	70,551	2,265,549
Medium Voltage	31	144	1,535
Low Voltage	38,558	70,407	2,264,014
Customers	84,169	134,882	4,299,800
Medium Voltage	31	495	11,503
Low Voltage	84,138	134,387	4,288,297
Electrical Buses	88,886	181,631	4,916,869
Electrical Nodes	168,005	375,334	9,868,205
Transmission substations	2	7	148
sub-transmission substations	8	31	632
distribution transformers	11,300	25,933	559,151
Line length (km)	1,921	4,576	116,837
Sub-transmission	27	167	4,128
Medium Voltage	966	2,302	64,460
Low Voltage (Secondaries)	928	2,107	48,249
Feeders	28	98	2,236

from multiple utilities around the U.S. The interval ranges for each is summarized on the right, along with the number of feeders in the validation set, which varies due to differences in the utility data availability/quality. Grades are assigned for each metric based on the prevalence of rare data. Good (G) corresponds to <5% rare, 5-10% is OK, and >10% is a flag to check (Chk) further. As described below (section IV.D.2) even a Chk is not grounds for immediate failure, but rather an indication to examine further. Even real utility systems compared in this report card produced Chk's for a few metrics.

Versions of these datasets are publicly available through both the Better Grids (https://bettergrids.org/) and DR Power (https://egriddata.org/) repositories by searching for the city names or looking under the NREL data collections.

IV. EXPERIENCES AND LESSONS LEARNED

This section describes a number of learning experiences from developing these large-scale distribution datasets. We hope that sharing these experiences and lessons learned will benefit both the growing community of synthetic electric systems developers and the users of the datasets by describing how challenges were overcome, a bit more about specific processes, and, in some cases, the tradeoffs in creating these data sets.

A. Pre-Processing Geospatial Input Data for Loads

The initial customer-level peak-load estimate utilizes geographical information: the location and dimension (height and footprint) of each building and the land or parcel use. We used a heterogeneous set of sources for this data including private vendors, public data available in OpenStreetMap, and local government websites, all of which required considerable preprocessing to clean up and harmonize.

1) Non-standard attribute encoding: One challenge is that the parcel attribute labels are not standard. Some use numerical codes, others use acronyms. Some distinguish between primary and secondary schools, others define a single category for any educational building. And so on. As a result, we developed a standardized parcel use label and then mapped each parcel label to a reference building type. For example, we grouped the reference buildings under 1) single-family, 2) multi-family, 3) stand-alone retail, 4) strip mall, 5) supermarket, 6) warehouse, 7) hotel, 8) education, 9) office, 10) restaurant, 11) outpatient healthcare, 12) hospital, and 13) industrial. All parcel labels were then mapped to these building types. This mapping allowed us to automate obtaining the peak load estimate for every building in any area independently of the initial labels in the data.

2) Managing mixed geometries: In addition to building-type-based load estimates, we also needed to assign a physical position for each building. Some data sources provide sufficient data to obtain the centroid of each individual building and locate the parcel polygon that encloses that centroid. This building-parcel matching not only provides the physical location, but also the building type information, since usage information is typically associated with the parcel and not included with the building geometry data. However, parcel

polygons are not always available, forcing us to use parcel centroids. To do so we mapped the closest parcel centroid to each building centroid. This may result in an incorrect location and hence allocation of category in the infrequent case when a building is located in the corner of a large parcel that is also next to a smaller parcel. However, empirical observation has shown that the same category buildings tend to group together; thus, reducing such allocation mistakes.

3) Handling outlier building sizes: After categorizing all buildings, we calculated the volume of each building assuming it is a rectangular prism, i.e., multiplying its footprint by its height. In the last step, the peak-load estimation of an individual building of each type (e.g. single-unit family) is obtained through a linear interpolation with respect to the same-type reference-building volumes. However, some buildings may be outside the size range of corresponding reference buildings. As a result, we added two dummy reference buildings to the interpolation. First, the so-called non-existent building with zero volume and zero peak-load consumption. Second, the largest building of each type in the area of study, for which the peak-load consumption can be obtained from expert sources, such as the building owners or the power utility. This maximum peak-load end point was key for the interpolation as it established a saturation and partially corrected the strong assumption of considering all buildings as constant height boxes. For larger, complex buildings/facilities (e.g. hospitals, colleges, or industries) the building height often varies, and this expert solicited maximum helps correct potential oversizing for the larger buildings.

B. *Adapting the Reference Network Model (RNM) for Creating U.S.-style distribution systems*

The RNM developed in the context of European distribution systems [12] was adjusted to design realistic US distribution networks. See [10] for a description of key modifications and algorithms needed to build US-style networks, such as

TABLE II: STATISTICAL VALIDATION RESULTS (PER FEEDER METRICS): SAF, GSO, AND SFO DATA SETS (V1.0)
KEY: TYP = TYPICAL, UNC = UNCOMMON, RAR = RARE, G=GOOD, CHK = CHECK, ϕ=Phase, OH=Overhead, Xfmr=transformer, len.=length, Cust=customer

Validation Metric	SAF Data Set Results				GSO Data Set Results				SFO Data Set Results				Utility Data Validation Regions		
	Typ	UnC	Rare	Grade	Typ	UnC	Rare	Grade	Typ	UnC	Rare	Grade	Typical	Uncommon	#Feeders
Dist. Xfrmr Tot. (MVA)	86%	14%	0%	G	84%	15%	1%	G	79%	20%	0%	G	[0+, 1.73], [4.94+, 31]	[1.73+,4.94], [31+,38.629]	5923
Tot. real load (kW)	79%	21%	0%	G	66%	31%	3%	G	52%	37%	11%	Chk	[4181+, 13793]	[577+,4181], [13793+, 17590]	1330
LV 1ϕ line len. (miles)	100%	0%	0%	G	100%	0%	0%	G	100%	0%	0%	G	[0+, 34.75]	[34.75+, 44.31]	57
LV 3ϕ line len. (miles)	100%	0%	0%	G	55%	34%	11%	Chk	85%	12%	3%	G	[0+, 1]	[1+, 2.135]	58
MV 1&2 ϕ line len. (mile)	100%	0%	0%	G	100%	0%	0%	G	99%	1%	0%	G	[0+, 35.36]	[35.36+, 124.62]	10632
MV 3ϕ line len. (miles)	100%	0%	0%	G	96%	4%	0%	G	97%	3%	0%	G	[0+, 20.84]	[20.84+, 45.6]	10149
MV OH 1&2 ϕ line ln (mi)	93%	7%	0%	G	97%	3%	0%	G	91%	9%	0%	G	[0+, 19.1]	[19.1+, 84.5]	10099
MV OH 3ϕ line ln (mile)	100%	0%	0%	G	96%	4%	0%	G	95%	4%	0%	G	[0+, 17.7]	[17.7+, 39.7]	9747
% of OH 1&2 ϕ lines	89%	7%	0.04	G	82%	18%	0%	G	81%	18%	1%	G	[0+, 0.23], [0.46+, 1]	[0.23+, 0.46]	9350
% of OH 3ϕ lines	86%	14%	0%	G	88%	12%	0%	G	89%	9%	2%	G	[0.4+, 1]	[0.18+, 0.4]	9492
# Cust.	100%	0%	0%	G	78%	20%	2%	G	76%	20%	4%	G	[94+, 2607]	[8+, 11837]	9734
Ratio of MV 1&2 ϕ line len. to num. Cust.	100%	0%	0%	G	95%	5%	0%	G	87%	12%	1%	G	[0+, 0.12]	[0.12+, 0.24]	9221
Ratio of MV 3ϕ line length to num. Cust.	64%	36%	0%	G	80%	20%	0%	G	70%	27%	3%	G	[0+, 0.09]	[0.09+, 0.77]	8556
# Fuses	100%	0%	0%	G	85%	15%	0%	G	82%	18%	0%	G	[4+, 187]	[187+, 281]	6013
# Reclosers	96%	4%	0%	G	97%	3%	0%	G	94%	6%	0%	G	[0+, 5]	[5+, 9]	6013
# Regulators	100%	0%	0%	G	100%	0%	0%	G	100%	0%	0%	G	[0+, 3]	[3+, 8]	11574
Sectionalizers	100%	0%	0%	G	100%	0%	0%	G	100%	0%	0%	G	[0+, 1]	[1+, 3]	5020
# Switches	82%	14%	4%	G	74%	20%	5%	OK	75%	19%	6%	OK	[3+, 392]	[392+, 635]	5020
# Cap. Banks	100%	0%	0%	G	100%	0%	0%	G	100%	0%	0%	G	[0+, 5]	[5+, 7]	11574
Avg. degree	100%	0%	0%	G	87%	13%	0%	G	89%	9%	1%	G	[1.9+, 2.06]	[1.6+, 1.9], [2.06+, 2.1]	5020
Char. path length. (miles)	96%	4%	0%	G	89%	11%	0%	G	87%	11%	1%	G	[12.4+, 95]	[2+, 12.4], [95+,134.39]	5020
Graph dia. (miles)	96%	4%	0%	G	90%	10%	0%	G	88%	11%	1%	G	[32+, 260]	[4+, 32], [260+, 371]	5020

changing the standard equipment used in US distribution networks regarding size and type of transformers, and power line characteristics (catalog of standard equipment), and managing criteria for underground vs. overhead power lines and transformers. And [17] explores the algorithms used for single, two- and three-phase feeder sections. However, there were few more esoteric challenges that had to be overcome and are presented here.

1) Overcoming the "wiggles": One of the first major issues when using RNM to build U.S. networks was the unrealistic wiggles—winding and bending—that appeared in the distribution networks. These are clearly visible in the left side of Figure 5 particularly in the sub-transmission network. We eventually tracked down the challenge to the nature of the street representation, which had a large number of roughly parallel lines representing sidewalks and other features, which allowed the graph heuristics to jump back and forth between very similar paths. We overcame this by simplifying the street maps to use only single lines and representing buildings as located at both sides of the street. The resulting clean final layout is shown on the right side of Figure 5.

2) Tree vs. star secondaries: The configuration of the secondaries was also an endless struggle and area of improvement in the model. The layout of real lines is often not modeled, which complicated verification. Initially, we adopted the simplest solution of using a star configuration (see Figure 6, left). However, by literally walking around in various parts of the U.S. and discussing with utilities, we realized that this was only sometimes used in practice by utilities, making it not suitable as the only approach.

Our second attempt used a tree configuration (Figure 6, center), where the buildings are connected to a pole in the street, and that pole is connected to the distribution transformer. This configuration looked more realistic to utility experts and relative to in-person observations, but was not universal since the star configuration is also found in practice. In our final approach (Figure 6, right) we opted for a hybrid configuration, where the star and the tree options are combined and seems to provide the most realistic secondary models.

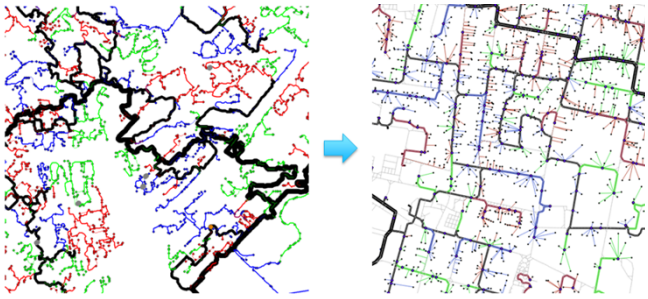

Figure 5: From wiggles (left) to final layout (right), here the three main voltage levels (low, medium and sub-transmission) are represented with ever thicker lines, and colors indicated phasing

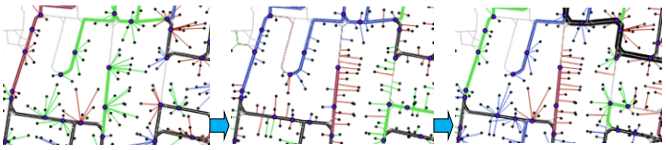

Figure 6: Evolution of the secondaries (L->R) from star configuration, to tree configuration and finally to a realistic hybrid configuration.

3) Adjusting heuristic parameters to better match statistics: The validation efforts were also critical to drive improvements in the model. As described in [13], the synthetic networks were validated from three points of view: i) statistically, ii) computationally, and iii) with input from utility and vendor experts. The expert feedback highlighted subtle structural characteristics of the networks, such as the configuration of the secondaries. However, the statistical validation made us rethink some of the heuristics used in the planning algorithms. For example, the comparison of the demand connected to each feeder in the utility validation data versus the synthetic networks showed discrepancies. Since customer demand is an exogenous input, this forced us to instead rethink how the demand of consumers was distributed among feeders. Figure 7, shows an example of a distribution network consisting of eight feeders. The modifications implemented changed the identification and definition of the area served by each feeder. In the final design of this case study, the full substation is finally fed with only four feeders, significantly increasing the demand of each feeder and hence better matching observed data.

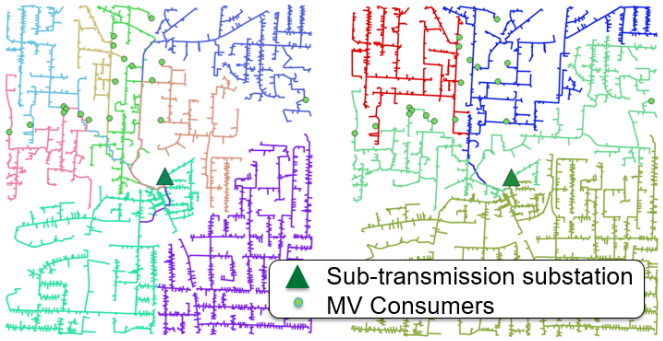

Figure 7: Evolution of the areas served by each substation: Substation with eight feeders (L), and the same loads and substation with four feeders (R)

4) Subregions for scalability: We encountered a number of challenges developing these very large networks. In particular, it was prohibitively slow to generate the large regions at once. Instead we divided the datasets into smaller subregions when creating the networks. This also allowed us to vary the design parameters and equipment catalogs by region to introduce additional diversity into the data sets. These separate regions were then connected through a high-voltage transmission network which connects to the sourcebus of each region. This greatly shortening the time for dataset creation.

C. The Need for Postprocessing

1) Adding Additional Technical Parameters Originally, we had hoped to be able to simply use the outputs of RNM-US

directly; however, we realized it was important to augment these to increase realism and provide sufficient electrical detail for some analyses. Specifically:

- **Adding substation internals:** The RNM-US substation representation as a single transformer is replaced with detailed substation internals including multiple transformers, fuses, switches, and breakers.
- **Replacing some switches with reclosers:** Adjusted to ensure no two reclosers are located in series.
- **Adding regulator set points:** Regulators are generally set to 1.03 p.u. output voltage, but based on power flow simulations, some feeders with low voltages are later adjusted to have feeder head voltage set to 1.05 p.u.
- **More accurate low voltage secondaries:** The representation of center-tap loads and triplex lines are modified to provide a more accurate representation of these elements.
- **Shift overlapping lines:** Line co-ordinates are adjusted slightly to avoid overlapping lines and nodes, primarily for visualization.
- **Adding fuse configurations:** Fuse limits are set to 100 A for medium voltage and 600 A for high voltage.
- **Adding disconnect switches to the start of long lines.**
- **Adding capacitor controls:** Capacitors were set to a default delay of 100 seconds, to switch on when the measured local voltage on a 120 V base is below 120.5 V, and switch off when it is above 125 V.
- **Sub-divide data into substations and feeders**: This allows users to more easily run individual substations and feeders rather than the entire region.

2) Tools for automation: With thousands of feeders to run and re-run it quickly became necessary to automate these post processing steps. To do we developed a pair of open source tools to do the heavy lifting. Specifically the Distribution Transformation Tool (DiTTo, https://github.com/NREL/ditto) enabled reading in data from a variety of formats, manupulating—such as the post-processing steps above, and then writing out the resulting set of files in multiple formats (e.g. OpenDSS, CYME, and an internal JSON for ease of use later). And the "Layerstack" workflow tool (https://github.com/Smart-DS/layerstack), enabled developing each of the post-processing steps as a separate "layer" that can be sequentially executed. A library of the Smart-DS specific layers is available opensource at https://github.com/Smart-DS/smartds-layerstack-library. These tools were also extensively used for collecting statistical summary data from thousands of actual distribution data files and for creating the rich scenario sets described below.

D. Validation and Refinement for Large Datasets

1) Automating Validation: Building on these same tools, we incorporated computing the statistical validation metric computation as a final step in the post-processing workflow. This greatly speed up validation feedback and hence refinement. Now rather than waiting to run an additional step, the comparison between the synthetic and real distribution system metrics could be quickly assessed speeding suggestions for modifying loads, network creation, or postprocessing.

We were also able to automate the operational validation by directly running power flow and other modeling assessments using OpenDSSDirect (https://github.com/dss-extensions/OpenDSSDirect.py). A key aspect of this effort was to produce voltage histograms for each region and voltage-distance plots for each feeder. These visual representations helped to quickly identify problem feeders and regions, as well as to provide insightful plots that could be included in the published dataset for analysis by other researchers.

2) Even Statistical Validation Requires Judgement: A key observation with our statistical validation efforts, was the realization that it was unrealistic (and perhaps unfair) to require that all of the synthetic datasets fully "pass" every metric, rather mismatches needed to be checked in order to make sure the discrepancy could be explained. We confirmed this approach by comparing individual actual utility service area data against the same combined, multi-utility reference used to validate the synthetic data. Even this real data had one to two metrics to "check."

3) Choice of validation metric: We also found it important to adjust the data used for statistical validation of total demand, specifically to replace historic demand with total capacity of distribution transformers connected to each feeder. This metric is not only more stable over time with changing customer habits, but also the transformer data is more readily available in the utility network planning models used for statistical comparisons, providing a much larger sample size and hence more valid comparisons.

E. Defining and Managing Many Large-Scale Scenarios

Our automation tools also enabled us to efficiently create scenarios and to match of timeseries to load. However, the diversity and scale of scenarios was large enough that it was not possible to provide separate datasets for every combination across technologies. Instead, each of the scenario dimensions and their corresponding levels are included as separate "layers" (the original motivation for the layerstack name) that can be mixed and matched to specify a desired combined scenario. Most of the scenarios are defined in terms of percentage of load points, which is different from customers since multi-family residential and multi-tenant commercial may have multiple customers at the same load point. Specifically, our generated core scenario sets include:

- Solar PV: 4 randomly created penetration levels each for rooftop and large (2MW) installations, the limits for each are considered separately (Table III). In addition, the inverter control settings are adjusted with penetration to be more aggressive at higher penetrations (Table III). The PV system sizes are based

on available customer land area (from parcel data) a seen in Table IV and Table V.

- Batteries: 2 penetration levels for each of small (behind the meter) and larger storage (Table VI and Table VII)
- Electric vehicles: 4 penetration levels for each of residential and commercial vehicles (Table VIII).
- Outage scenarios: 4 levels of severity (Table IX)
- Controllable loads: 3 levels of loads that are assigned to be remotely controllable for each of residential and larger customers (Table X)
- Smart Meter visibility: 4 penetrations of smart meter deployments (Table XI)

TABLE III: SOLAR SCENARIOS (ROOFTOP AND LARGE CAN BE SEPARATE)

Scenario	% Load Pts. with "Rooftop" Solar	%Feeders with one/two Large PV	Max. Solar Roof/Large (%Feeder Peak)	Inverter Controls for additional PV
(Base)	0%	0%	-	-
Low	15%	None	15% / 0%	1547-2013
Medium	35%	50% / 0%	75% / 33%	PF=0.95 absorbing
High	65%	100% / 75%	150% / 80%	1547-2018 Cat. A: half: V/Var-only, half: V/var & V/watt
Extreme	85%	100% / 75%	No limit / 100%	1547-2018 Cat. B: V/var & V/watt

TABLE IV: RESIDENTIAL SOLAR INSTALLATION SIZES

Land Area (ft^2)	Residential Installation Size (kW)
<75	3
75–300	5
>300	8

TABLE V: COMMERCIAL SOLAR INSTALLATION SIZES

Land Area (ft2)	Commercial Installation Size (kW)
<100	3
100–300	6
300–600	8
600–1000	40
1000–2000	100
>2000	300

TABLE VI: SMALL BATTERY INSTALLATION SIZES

Installed PV capacity (kW)	Installed Battery Size (kW)
< 4	4
4–10	8
10–150	25
> 150	100

TABLE VII: BATTERY SCENARIOS

Scenario	% Load Points with Small Bat.	% of Substations with One/two Large Bat.
(Base)	0%	0%
Low	5%	50% / 0%
High	35%	100% / 75%

TABLE VIII: ELECTRIC VEHICLE SCENARIOS (EACH DIMENSION CAN BE SELECTED SEPARATELY)

Scenario	Residential Load Points with Level-2 Chargers	Commercial Load Points with Level-2 Chargers	Feeders with DC Fast Chargers
(Base)	0%	0%	0%
Low	5%: 1 car	5%	1%
Medium	30%: 1 car	30%	5%
High	60%: 1 car 15%: second car	75%	10%
Extreme	75%: 1 car 45%: second cars	100%	25%

TABLE IX: OUTAGE SCENARIOS

Scenario	Outages
(Base)	None
Low	1 line IDed per feeder
Medium	3 lines IDed per feeder
High	2% of lines, randomized by region
Extreme	20% of lines, randomized by region

TABLE X: CONTROLLABLE LOADS (EACH DIMENSION CAN BE SELECTED SEPARATELY)

Scenario	Residential & Small Commercial	Large commercial (> 200kW)
(Base)	0%	0%
Low	5%	15%
Medium	30%	50%
High	75%	100%

TABLE XI: SMART METERS (ADVANCED METERING INFRASTRUCTURE, AMI)

Scenario	Loads with AMI
(Base)	None
Low	5%
Medium	15%
High	75%
Extreme	100%

Some of these scenario types (e.g. solar), can be directly captured in distribution modeling formats (all data sets available in both OpenDSS and CYME formats), while others are not as well defined in standard tools. Moreover the full factorial combination of every scenario dimension would result in over 2.5 million combinations. As a result, we opted to provide a combination of model-based scenarios and "placements." The placements are additional files listing which customers/nodes have the scenario attribute that then can be mixed and matched as desired by the user to produce an arbitrary combination from the full range of combinations without the combinatorial challenge of offering unique data sets for each. These "placements" form the core of the mix and match layers and can be referenced in future research for consistency across studies.

V. Discussion and Conclusion

This paper has described our experiences in developing large-scale synthetic U.S.-style distribution data sets. Particular emphasis is placed on implementation details that are normally not described in detail in the literature, with the hope that lessons learned from our experiences can be useful for both our users and others developing synthetic grid models.

Specifically, we found that while developing new algorithms and modeling approaches was critical to create large-scale realistic test systems for U.S. systems, considerable effort was also required to simply manage all of the data for such large systems. We have hence described various automation tactics across the entire dataset workflow and highlighted a few practical new ideas such as the use of mix-and-match placements to manage the large number of scenario combinations.

The resulting open access datasets provide realistic scales, and complete model details to enable a wide range of next generation algorithmic research. Example applications include evaluating scalability, performance, and/or impact of: novel distribution automation approaches, advanced distributed ACOPF algorithms, distributed energy resource management systems (DERMS), electrified transportation, DERs providing grid services, distribution system operator (DSO) performance, distribution market designs, and many other applications.

Acknowledgement

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 No. DE-AC36-08GO28308. Funding provided by the Advanced Research Projects Agency–Energy. 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] F. E. Postigo Marcos *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, Nov. 2017, doi: 10.3390/en10111896.

[2] S. Babaeinejadsarookolaee *et al.*, "The Power Grid Library for Benchmarking AC Optimal Power Flow Algorithms," *ArXiv190802788 Math*, Aug. 2019, Accessed: Apr. 19, 2020. [Online]. Available: http://arxiv.org/abs/1908.02788.

[3] A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, and T. J. Overbye, "Statistical Considerations in the Creation of Realistic Synthetic Power Grids for Geomagnetic Disturbance Studies," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1502–1510, Mar. 2017, doi: 10.1109/TPWRS.2016.2586460.

[4] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," in *2018 IEEE Texas Power and Energy Conference (TPEC)*, Feb. 2018, pp. 1–6, doi: 10.1109/TPEC.2018.8312059.

[5] K. M. Gegner, A. B. Birchfield, Ti Xu, K. S. Shetye, and T. J. Overbye, "A methodology for the creation of geographically realistic synthetic power flow models," in *2016 IEEE Power and Energy Conference at Illinois (PECI)*, Feb. 2016, pp. 1–6, doi: 10.1109/PECI.2016.7459256.

[6] 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, doi: 10.1109/TPWRS.2016.2616385.

[7] W. H. Kersting, "Radial distribution test feeders," in *IEEE Power Engineering Society Winter Meeting*, 2001, vol. 2, pp. 908–912 vol.2, doi: 10.1109/PESW.2001.916993.

[8] K. P. Schneider, Y. Chen, D. Engle, and D. Chassin, "Taxonomy of North American radial distribution feeders," in *Proceedings of the Power & Energy Society General Meeting*, Calgary, AB, Jul. 2009, doi: 10.1109/PES.2009.5275900.

[9] B. Palmintier and B.-M. Hodge, "Beyond the Feeder: Large-scale synthetic distribution systems for next-generation algorithms and analysis," presented at the IEEE PES General Meeting, Portland, OR, Aug. 07, 2018, [Online]. Available: https://www.researchgate.net/publication/336238973_Beyond_the_Feeder_Large-scale_synthetic_distribution_systems_for_next-generation_algorithms_and_analysis.

[10] C. Mateo *et al.*, "Building Large-Scale U.S. Synthetic Distribution Network Models," Madrid, Spain, Working Paper IIT-19-004A, Feb. 2019. [Online]. Available: https://www.iit.comillas.edu/publicacion/mostrar_publicacion_working_paper.php.en?id=352.

[11] 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 (United States), NREL/TP-5500-46861, Feb. 2011. doi: 10.2172/1009264.

[12] C. Mateo Domingo, T. Gomez San Roman, Á. Sanchez-Miralles, J. P. Peco Gonzalez, and A. Candela 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, doi: 10.1109/TPWRS.2010.2052077.

[13] V. Krishnan *et al.*, "Validation of Synthetic U.S. Electric Power Distribution System Data Sets," *IEEE Trans. Smart Grid*, 2020, doi: 10.1109/TSG.2020.2981077.

[14] 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," National Renewable Energy Lab. (NREL), Golden, CO (United States), NREL/TP-5500-68670, Dec. 2017. doi: 10.2172/1414819.

[15] E. Hale *et al.*, "The Demand-Side Grid (dsgrid) Model Documentation," National Renewable Energy Lab. (NREL), Golden, CO (United States), NREL/TP-6A20-71492, Aug. 2018. doi: 10.2172/1465659.

[16] T. Elgindy *et al.*, "Synthetic San Francisco Bay Area (SFO) Electrical Distribution System Technical Description," National Renewable Energy Laboratory, Golden, CO, NREL Technical Report NREL/TP-5D00-74245, In preparation 2020.

[17] 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, p. 106042, Sep. 2020, doi: 10.1016/j.ijepes.2020.106042.