A Real-World Test Distribution System with Appliance-Level Load Data for Demand Response and Transactive Energy Studies

FERNANDO B. DOS REIS1, (Member, IEEE), REINALDO TONKOSKI2, (Senior Member, IEEE), BISHNU BHATTARAI1, (Senior Member, IEEE) and TIMOTHY M. HANSEN2, (Senior Member, IEEE)

1Electricity Infrastructure and Buildings Division, Pacific Northwest National Laboratory, USA
2Electrical Engineering and Computer Science Department, South Dakota State University, Brookings, SD 57007 USA

Corresponding author: Fernando B. dos Reis (e-mail: fernando.beretadosreis@pnnl.gov).

This work is supported by the National Science Foundation (NSF) under grant number ECCS-1608722, the South Dakota State University Ph.D. Fund, and South Dakota Board of Regents (SDBoR). This research made use of the “Roaring Thunder” cluster at South Dakota State University funded under National Science Foundation grant number CNS-01726946.

ABSTRACT

Research on demand response and transactive energy systems often require granular, appliance-level demand data. However, there is no existing test system with such appliance-level data with proper temporospatial diversity in a realistic distribution system. This paper develops a 240-node real distribution test system with appliance-level demand data for responsive loads. The residential appliance-level demand data are derived from smart meters connected to 1,120 homes in a real distribution system from Iowa State, hereafter called Midwest 240-Node test distribution system. A queueing load model was used to derive the appliance-level data from the smart meter data. The Midwest 240-Node test distribution system provides granular appliance-level information for all homes in the distribution system (i.e., individual appliances that constitute the home load), and the aggregate of all customer load emulates the actual smart meter data. The performance of the Midwest 240-Node test distribution system is evaluated by comparing the aggregated appliance-level demand with the actual measured smart meter data from the utility. The one-year appliance data has a mean absolute percentage error of 2.58% compared to the measured smart meter data. The test system is modeled in OpenDSS and GridLAB-D and is openly available to researchers to enable demand response and transactive energy studies with active end-users.

INDEX TERMS Demand response, distributed energy management, home energy management, queueing load model, synthetic distribution test system, transactive energy.

I. INTRODUCTION

Given the growing variability in generation from the participation of non-dispatchable resources (e.g., wind and solar), there is a need for increased operational flexibility for the future environmentally friendly, economical, and secure power system [1]–[4]. Demand response (DR) is one such source of flexibility, and according to [5], [6], is a main component of the smart grid. DR encourages consumers to change their demand concerning power system conditions.

Residential loads represent approximately 38% of the total energy consumption in the U.S. [7]. Residential DR can provide the needed operational flexibility, and the major resulting benefits are: (a) participant financial benefits; (b) market-wide financial benefits; (c) reliability benefits; and (d) market performance benefits [4], [8]–[10]. Residential DR makes system-wide changes that require tens of thousands of buildings, each with many individual electric energy devices,
to be controlled [11]. A generalized form of DR that considers or coordinates both supply and demand is commonly referred to as transactive energy [12]–[14].

There is a missing link in the research community between the availability of aggregate power system demand, the individual customer demand that composes it, and the location of such demand on distribution system networks. Multiple home energy management system studies neglect their impact on the power system [15]–[18]. In [15], [18], [19], residential DR optimization models for scheduling individual customer appliances are presented. The work in [15] has the customers as price takers utilizing actual real-time pricing information from an Illinois power company. In [18], coordination and market interaction is considered utilizing the emulation of an actual U.S. electricity market. The paper utilizes a hierarchical controller framework bidding strategy for demand reduction events considering the consumer preferences. Differently than [15]–[17], a considerable effort is presented to evaluate the change in locational marginal price given the change in demand [18]. The aggregation of customers to perform the optimization is utilized by [16], [17]. The work in [16] is similar to [15] with the difference of utilizing aggregation. In [17] the focus is on multi-objective optimization trade-off between customer financial benefits and customer discomfort. Customer comfort is directly considered in the optimization in [17], [18], [20]. However, the studies presented in [15]–[18] do not have a residential customer demand that aggregates to a known region of the power system. A combined power system and home energy management test system must emulate the behavior of a real distribution system with individual customer loads that aggregate to the known system load. Existing distribution test systems do not have real time-series load data, except for the IEEE European LV [21], but this is only for a single day. By linking individual customer loads to the system load, calculation and analysis of system-level impacts of residential DR is enabled. Such analysis allows studies to more accurately demonstrate the flexibility and impacts of DR on power system operations (e.g., electricity markets, and reduction in renewable energy curtailment).

In [22] the availability of a power system test case with granular device level data that accurately represents the behavior of a real system (e.g., electric grid, consumer, and responsive end-use loads) can be an enabler for the design and analysis of new and scalable approaches for the integration of distributed energy resources. A synthetic distribution test system using street maps, equipment catalogs, and building expected behavior was created to test demand-side management algorithms with considerable number of distributed resources present in distribution systems [23]. The paper [22] continues the work of [24]–[28] for the creation and validation of synthetic transmission systems and [29]–[32] for the creation and validation of synthetic distribution systems with access to utility data. The work presented in [22] focuses on the U.S. distribution systems, creating and validating synthetic distribution systems of up to 10 million electric nodes. The authors from [22] are researchers at National Renewable Energy Laboratory (NREL) and have access to real utility data, which is considered Critical Energy/Electric Infrastructure Information (CEII) and is not available to the general research community. Their approach enables the validation of flexible synthetic distribution systems, but the methods still require unavailable CEII. Additionally, the work in [22] validates three large-scale synthetic test systems, with statistical quantification to infer how realistic the networks are compared to real data. Differently than the work from [22], this paper creates time-series synthetic load data at the granular-level from aggregated smart meter data on a real distribution network.

The Iowa State distribution system test feeder in [33] made available a real distribution network from the U.S. Midwest region with one-year smart meter node-level load data for 2017. This unique test feeder combines a real utility distribution system network model with corresponding field measurements that are publicly available. To maintain individual consumer privacy, the available data is aggregated to node-level and is provided in an hourly resolution. Preserving consumer privacy is a necessity, however, granular-level residential load data to test transactive energy studies is also required. Thus in this paper, the nodal load data is first divided into 1,120 homes across 193 load nodes over three feeders. The home data is further divided into appliance-level data using the queuing load model from [34]. The one-year mean absolute percentage error (MAPE) between the real smart meter data and granular-level synthetic data generated for the Midwest 240-Node test distribution system is 2.58%. The main contributions of this paper are:

(a) the verification of queuing load models for generating granular synthetic residential device-level load data from nodal real-world smart meter data; and
(b) the development and validation of the open-source Midwest 240-Node test distribution system, which is the first test system with appliance-level data with temporal diversity in a realistic distribution system.

The open-source Midwest 240-Node test distribution system is provided in both OpenDSS and GridLAB-D and was validated with a maximum voltage magnitude error less than 10−3%. This test feeder will enable researchers to perform granular-level smart grid and transactive energy studies, and measure the system-level impacts.

The remainder of this paper is organized as follows: Section II is an overview of the publicly available Iowa State distribution system test feeder. The queuing load model for generating the granular-level synthetic load data from the node-level smart meter data is described in Section III. In Section IV, the Midwest 240-Node test distribution system is validated in regards to both the load mismatch (real nodal smart meter data vs. synthetic granular-level data) and power flow impact (OpenDSS vs. GridLAB-D). Section V presents the main conclusions of this study. Finally, the appendix details the conditioning of missing smart meter data from the.
published Iowa State distribution system test feeder for use in the Midwest 240-Node test distribution system.

II. DESCRIBING THE IOWA STATE DISTRIBUTION SYSTEM TEST FEEDER

Power system test cases, including distribution test systems, are derived from the general characteristics of real networks. The Iowa State distribution system test feeder is a publicly available distribution network from the Midwest U.S. [33] developed in OpenDSS. The test system has 240 primary network nodes and 23 miles of primary feeder conductor. In addition to the real network data, one-year smart meter measurements at the node-level was also provided.

![FIGURE 1. One line diagram of the test system. Adapted from [33].](image)

The Iowa State distribution system test feeder is presented in Fig. 1 as a radial distribution system consisting of three feeders [33]. The feeders are labeled as S, M, and L, referring to the relative size of the feeders as small, medium, and large, respectively. A 10 MVA delta-wye step-down 69/13.8 kV substation transformer supplies power for the three feeders. The substation transformer has a tap-changer mechanism that consists of three independent single-phase tap changers. Feeders M and L have shunt capacitor banks for voltage regulation. The utility has a strategy to switch on capacitor banks in normal operation to provide reactive power support. Iowa State distribution system test feeder has nine circuit breakers at the illustrated locations in Fig. 1 that are used for protection and reconfiguration. Six of the circuit breakers are normally closed, and three are normally open. All standard electric components in the Iowa State distribution system test feeder are modeled, such as overhead lines, underground cables, substation transformers with load tap changers, line switches, capacitor banks, and secondary distribution transformers.

The Iowa State distribution system test feeder has 1,120 homes, each with an installed smart meter [33]. There are 193 system load nodes with 15 on Feeder S, 44 on Feeder M, and 134 on Feeder L, each with a unique numeric number from 0 to 192. The assigned number for the load node follows the order from the provided files in [35] and is read in the following order: S, M, and L. The homes are connected to the primary network nodes via secondary distribution transformers, demonstrated in Fig. 1. The load data is measured using smart meters for the year 2017 in an hourly resolution (in kWh) by approximating the hourly energy consumption under the assumption that the customer demand is constant in each one-hour time interval [33]. To model reactive power for the load nodes, a power factor is randomly selected in the range of 0.9–0.95 for each homes [35]. The power factor and reactive power of each customer is calculated and aggregated for the customers in the same load node.

Although 1,120 homes are known to be on the network, the provided load data in [33] is aggregated at the node level to protect the privacy of individual customers. Additionally, it is unknown if any customers have distributed generation, such as solar photovoltaic.

III. GENERATING GRANULAR-LEVEL SYNTHETIC LOAD DATA

A. OVERVIEW

The real customer demand from smart meter measurements [35] are the aggregation of customers at a given load node with hourly resolution. Power system studies such as home energy management systems, distributed energy management, DR, and transactive energy require high-resolution individual customer load (i.e., the knowledge of appliances that compose the demand of each customer/home). To utilize the data provided in the Iowa State distribution system test feeder for such studies and taking advantage of the real customer demand data, in this section (a) the provided nodal load data is analyzed, and time periods with erroneous smart meter data are statistically replaced (more information in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3120923, IEEE Access
the gamma parameters $k$ and $\theta$ are computed with $k = \mu^2 / \sigma^2$ and $\theta = \sigma^2 / \mu$. For the granular-level data, the appliance set for each home is generated by sampling two distinct gamma distributions, one for the power rating of the appliances and the other for the duration of the appliances. Gamma distributions are continuous probability distributions in the positive real number set defined by two parameters (i.e., shape $k$ and scale $\theta$). The mean of a gamma distribution is $\mathbb{E}[X] = k\theta$, and the variance is $\text{Var}(X) = k\theta^2$. Thus, by defining the mean $\mu$ and standard deviation $\sigma$, the gamma parameters $k$ and $\theta$ are computed with $k = \sigma^2 / \mu^2$ and $\theta = \sigma^2 / \mu$. For the granular-level data, the appliance set is generated with gamma distribution parameters as power (W) $\mu = 500$ and $\sigma = 100$, and appliance duration (hour) $\mu = 0.5$ and $\sigma = 0.25$, as utilized in [34].

D. SYNTHETIC QUEUEING LOAD MODEL

Queueing models are defined by the probability distribution of inter-arrival times $T$ (i.e., appliance inter-arrival times), probability distribution of service times $X$, number of servers $C$ (i.e., power supply capacity), queue capacity, size of the population, and a service discipline. Furthermore, the characteristics can be constant or time-dependent (e.g., inter-arrival times as a function of time, as illustrated in Fig. 2). The queueing load models in [34] make three assumptions: the queue length is infinite (i.e., no loss of appliances arriving at the system); the population is infinite (i.e., arrival process is not dependent on the appliances currently present in the system); and the service policy is first come first served. Given the assumptions, the queueing load models are described with the simplified Kendall notation, i.e., $T/X/C$.

The synthetic queueing load models combine a top-down bottom-up approach, where the expected load of a customer ($l(t)$) is used as the input for computing statistical time-varying customer appliance arrival rates. The appliances are modeled as generic blocks of energy as in [34]. Fig. 2 presents a summary of the process for generating the synthetic queueing load for one of the load nodes. Thus, the output of the synthetic queueing load models have granular-level appliance data that constitutes the demand for each customer at each load node for the Midwest 240-Node test distribution system. The numbers from 1 to 4 with the arrows on a light green background are the steps for generating the load. Step 1 creates reference curves for all customers that constitute a given aggregate load node demand. In Step 2, the queueing load models run as independent processes for each customer (using the reference curve from Step 1). The queue outputs the generated appliances for each customer over the given time period in Step 3. Step 4 aggregates the appliance loads for all the customers and statistically approximates the original load node demand.

Three queueing load models are presented in [34], i.e., the $M_t/G/\infty$, $M_t/G/C$, and $M_t/G/C_t$. The models have a time-varying probability distribution of inter-arrival times ($M_t$) and the probability distribution of service times is general ($G$). However, each queueing load model has a distinct power supply capacity, being infinite ($\infty$), constant ($C$), and time-varying ($C_t$), respectively. Given the natural random characteristic of the queueing models with the probability distribution of inter-arrival times, it is expected that the larger the number of customers being generated, the smaller the deviation will be from the reference curve for a given load node. The formulation for the queueing load models and further explanation are presented in [34]. Loads that need to be treated separately from the queueing arriving arrivals can be simply subtracted from $l(t)$, as shown in [34] (e.g., thermal loads, electric vehicles).

E. PARAMETERS FOR THE QUEUEING LOAD MODEL

According to U.S. Energy Information Administration 2015 Residential Energy Consumption Survey [37], homes from the Midwest region have an average yearly consumption of 9,567 kWh. Residential energy consumption is climate dependent. The climate dependence characteristic is demonstrated in [33]. To minimize the climate dependency for defining the number of homes by nodal load, the month of May is selected as it is neither winter nor summer, and thus will minimize the impacts of customer variance due to thermal loads. Using the month of May for selecting the number of homes, without considering the periods of strange behavior (i.e., the period from May 1 to May 25 is utilized, further explained in the appendix), and assuming an even daily energy consumption from the yearly average consump-
FIGURE 2. Summary of the synthetic queueing load model used to generate the granular-level data for each home on the Midwest 240-Node test distribution system. At each load node, the node-level load is split into a per-home load reference curve, denoted by ‘1.’ Each home independently generates the granular-level appliance data using the synthetic queueing load model (‘2’ and ‘3’). Lastly, denoted by ‘4,’ the aggregated load from appliances from all homes on the load node will statistically represent the node-level reference curve.

The calculated definition of a Midwest home, results in the calculated number of homes being 1,187 (i.e., the average daily energy use from May 1 to 25 per load node divided by the average daily energy use from the Energy Information Administration).

The actual number of customers in the system is 1,120, thus there are 67 extra homes from the calculated that have to be removed to match the known number of customers. The queueing load model presented in Section III-D better approximate the desired load curve with height reference energy values as demonstrated in [34] (i.e., height $l(t)$). The reference curve for a load node curve $l(t)_n$ is computed with $l(t)_n = \text{load}_n / NH_n$, where $\text{load}_n$ is the node load, $NH_n$ is the number of homes by node, and the subscript $n$ represents the individual load node index.

To improve the load node reference curve (i.e., $l(t)_n$) and remove the 67 extra homes the algorithm presented in Fig. 3 is utilized. The algorithm requires the information of the node load (i.e., $\text{load}_n$), number of homes by node (i.e., $NH_n$), and average yearly consumption $\varphi$, excluding the periods of strange behavior for $\text{load}_n$ and $\varphi$ (as described in the appendix).

The algorithm removes one home at a time, giving priority to homes with lower energy use if two conditions are satisfied. First, the node in question must have more than one home (i.e., a load node cannot have less than one home). Second, the resulting energy consumption at the node will not surpass 1.5 times the $\varphi$. The value of 1.5 was chosen to avoid deviating too much from the average yearly consumption.

The resulting number of homes by load node is presented in Fig. 4. Presenting the number of homes before and after running Algorithm 1 that removes the extra homes.

The synthetic queueing load model $M_t/G/C_t$ was chosen to generate data for presenting as it produced a lower deviation from the smart meter data than the other queueing models. The time-varying power supply capacity ($C_t$) is never allowed to be smaller than 1,500 W (i.e., the input to the queueing load model is never less than 1.5 kW), and is created by adding a gain of 2 to the customer expected load (i.e., $C_t$ is $2 \times$ the reference curve). These were selected so statistically there is always the possibility of serving multiple
FIGURE 3. Algorithm that removes the extra homes from Midwest 240-Node test distribution system.

Algorithm 1: Removes the extra homes.
1) $load_n \leftarrow$ load at every node $n$
2) $NH_n \leftarrow$ number of homes per node
3) $\varphi \leftarrow$ average yearly home energy consumption
4) while $\sum_{n=0}^{192} NH_n > 1,120$ do
5) \[ v_n = \sum_{n=0}^{192} load_n/NH_n \leftarrow \text{home energy} \]
6) $i \leftarrow$ index order from smaller to larger $v_n$
7) for $(j = 0; j \leq 192; j = j + 1)$ do
8) \[ N_{H_{[j]}} > 1 \& \sum_{NH_{[j]-1}}^{load_{[j]}} \frac{1}{1.5 \times \varphi} \]
9) \[ N_{H_{[j]}} \leftarrow N_{H_{[j]-1}} - 1 \]
10) break

FIGURE 4. Number of homes by load index considering the period of May 1 to May 25 and the updated number of homes (i.e., the output from the algorithm).

appliances, as well as suppressing unrealistic and infeasible load peaks [34].

IV. VALIDATION

A. OVERVIEW

This section presents the validation of the proposed approach for the generated granular-level synthetic load compared to the original node-level smart meter data. The synthetic load generated with the queueing load model utilizes the nodal load smart meter real data for the year of 2017 divided by the expected number of homes of that node, making the expected load of a customer ($l(t)$). With that, the queueing load model is run generating the arrival of appliances for each customer to generate the load per customer and the appliances that compose the customer load.

B. SMART METER VS. SYNTHETIC LOAD

The queueing model is a random process of arrival of appliances and, as such, the generated load will be different from the smart meter data. In this section, the periods of strange smart meter behavior presented in Section III-B and further explored in the appendix are not considered. A metric utilized for evaluating the distance or error from the smart meter load and synthetic load is the MAPE,

\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|, \tag{1}
\]

where $A_t$ is the smart meter data and $F_t$ is the generated synthetic load. The subscript $t$ is the discrete time interval in an hourly resolution. Thus, the synthetic load is converted to hourly consumption to enable the comparison (although the appliances are provided at a smaller time resolution).

From [34] it is known that one of the characteristics to reduce the difference from the reference curve is the number of aggregated customers. Thus, with a larger number of customers being aggregated to a node load it is expected there is a smaller error. Fig. 5 demonstrates this characteristic. The nodes and their respective MAPE over the year are plotted with the load nodes ordered from smallest to largest with respect to their number of customers. As it can be observed, there is a negative correlation between the number of homes and MAPE.

FIGURE 5. Year MAPE by load node in relation to the number of homes.

The last three nodes in Fig. 5 present peculiar behavior as they are the nodes with the most homes, but appear to have an increasing MAPE. Table 1 presents more variables to assist in explaining this behavior, i.e., load node numeric identifier (Node), number of homes ($NH$), minimum customer load (min), median customer load (median), customer energy (energy), and load node MAPE. The table presents the last six load nodes from Fig. 5, having a line to separate Nodes 52, 51, and 56 (i.e., more $NH$, lower MAPE) from Nodes 120, 40, and 15 (i.e., more $NH$, larger MAPE). Other characteristics that increase the difference from synthetic to a reference are the low values of the reference curve being summarized in Table 1 with minimum power, median power, and energy. Low values of the reference curve are problematic given that some periods are not likely to have appliances arriving and/or large inter-arrival periods [34]. Node 120 is like Node 51 in minimum load and energy; however, the low
median increased the MAPE. Node 40 is like Node 52 in energy; however, the higher minimum and median lowered the MAPE. Node 15 is like Node 56 in energy; however, the lower minimum and median increased the MAPE.

TABLE 1. Explaining the number of homes deviation from smaller MAPE.

| Node | NH | min (W) | median (W) | energy (MWh) | MAPE (%) |
|------|----|---------|------------|--------------|----------|
| 52   | 21 | 156.19  | 681.90     | 12.99        | 14.72    |
| 51   | 40 | 238.05  | 1567.32    | 13.09        | 6.50     |
| 56   | 42 | 433.98  | 1063.80    | 9.26         | 6.07     |
| 70   | 48 | 214.58  | 836.79     | 13.24        | 7.12     |
| 40   | 58 | 413.79  | 1368.27    | 11.39        | 12.09    |
| 15   | 60 | 357.00  | 677.66     | 8.67         | 15.73    |

The knowledge of the MAPE by explored node for the year in Fig. 5 is further explored graphically in Fig. 6, where MAPE is computed per-day for all days and all nodes. Furthermore, the resulting day and node MAPE is ordered in decreasing order. Fig. 6 graphically presents five days, giving an insight into the daily behavior. Each of the five plots has a title containing the load node number, followed by the percentile of the day/node in order of MAPE. Thus, 0% is the worst performing (0th percentile) and 100% is the best, with 5% and 10% being on the lower end, and 50% as the median. The comparison shows that even the poorer performing nodes (e.g., 10%) are still quite well-performing.

As expected and demonstrated in Figs. 5 and 6, aggregating more customers reduces the difference between the smart meter load and the synthetic load. In a similar fashion to the previously described figure, Fig. 7 presents the worst, median, and best load days in four levels of aggregation. The four levels of aggregation are each row as the three feeders of the system and the total distribution system, i.e., Feeder S, Feeder M, Feeder L, and system (i.e., entire Midwest 240-Node test distribution system). Each plot in Fig. 7 has a title containing that particular level of aggregation day MAPE. The smallest level of aggregation is Feeder S, that for the worse day is 7.92% MAPE, i.e., no other day in any other presented level of aggregation will perform worse. Feeders S, M, and L have a clear reduction of MAPE from smaller to larger.

The levels of aggregation are ordered from smaller to larger in Fig. 7. However, Feeder L has lower MAPE for the worse and best day than the complete distribution system. This occurred given that we are presenting the MAPE for the day being presented. Table 2 presents the levels of aggregation in relation to the number of customers and the yearly MAPE. Demonstrating that increasing the aggregation will reduce the MAPE. Even though the MAPE gets better with increased aggregation, even the worst performing days and load nodes still perform quite well.

TABLE 2. MAPE for the year of 2017 in four levels of aggregation, i.e., Feeder S, Feeder M, Feeder L, and system.

| Feeder | NH   | MAPE (%) |
|--------|------|----------|
| S      | 76   | 6.4617   |
| M      | 370  | 3.8090   |
| L      | 674  | 2.5864   |
| System | 1.120| 2.5828   |

C. POWER LOAD COMPARISON

The different input loads impact for the Midwest 240-Node test distribution system power flow are presented in this
FIGURE 7. The worse, median, and best load days are presented in four levels of aggregation, i.e., Feeder S, Feeder M, Feeder L, and system (i.e., entire Midwest 240-Node test distribution system). On the top of every plot is the day MAPE.

FIGURE 8. Distribution of voltage magnitudes for smart meter and synthetic load data for Midwest 240-Node test distribution system for one-year. The labels 1 ST, 2 FS, 3 FM, and 4 FL, refer to the substation transformer primary side, Node 10 from Feeder S, capacitor node Feeder M, and capacitor node Feeder L, respectively. The median is shown with a dashed line, and the upper and lower quartiles with the dotted lines.

Section IV-B demonstrated that smart meter and synthetic load data are not exactly the same, but the synthetic load follows the behavior of the smart meter load data. The impact on the power flow is presented by demonstrating the voltage behavior at four points of Midwest 240-Node test distribution system showing the annual distribution of voltage magnitudes in p.u. for phases A, B, and C as presented in Fig. 8 (8,760 voltage magnitude samples for every half violin plot; thus a total of 210,240 voltage magnitude samples for the four nodes and two load types using one-hour time resolution over one-year). This type of plot is like a box plot, but with the (rotated) kernel density plot on each side. The thickness (or density) represents how often each voltage magnitude occurred.

The node being presented in Fig. 8 are labeled 1 ST, 2 FS, 3 FM, and 4 FL, which refer to the substation transformer primary side, node 10 from Feeder S, capacitor node Feeder M, and capacitor node Feeder L, respectively. The location of the presented nodes are identifiable in Fig. 1. The three phase nodes were chosen empirically with the intent of demonstrating the voltage at the substation and within each of the feeders. The violin plots in Fig. 8 are split in half, with smart meter data on the left side and the synthetic data on the right. Each half of the split violin plots have similar shapes, and the median and quartiles are near each other. Thus, the power flow studies with the generated synthetic load approximate the behavior of the smart meter load.

D. CREATED GRIDLAB-D MODEL

The GridLAB-D simulation software was created for distribution systems. The core of GridLAB-D has an advanced algorithm that simultaneously coordinates the state of millions of independent devices, each of which is described by multiple differential equations. GridLAB-D examines in detail the interplay of each part of a distribution system with every other, and incorporates an extensive suite of tools to build and manage studies and analyze results, e.g., agent-
based and information-based modeling tools that allow users to create detailed models of how new end-use technologies, distributed energy resources, distribution automation, and retail markets interact and evolve over time. Thus GridLAB-D as a tool is of interest for multiple power system studies, especially for smart grids, smart cities, demand response, and home energy management systems [38].

The OpenDSS model available from [35] was converted to GridLAB-D making use of the Python packages DiTTo [39] and glm [40]. The DiTTo package made an initial conversion file for GridLAB-D, however it does not consider the split phase structure and the impedance to which the distribution system is connected to the main grid. The package “glm” was used to address the split phase structure and some unity mismatches. The impedance from the swing node to the distribution system was computed as presented in [41]. Table 3 presents the comparison of voltage and current from OpenDSS and GridLAB-D, for a single power flow solution. The single power flow solution is the original available from [35]. The voltage magnitude comparison is performed for all nodes; Table 3 presents only the worse for each of the phases. The percentage-wise maximum error observed for voltage magnitude is below $9 \times 10^{-4} \%$. The current magnitude comparison was performed for the lines and transformer primaries, however currents below 0.1 A on OpenDSS are not considered; Table 3 presents only the worse for each of the phases. The percentage-wise maximum error observed on the considered current magnitude is below 0.04%. The GridLAB-D model with the synthetic load data is made publicly available at [42] and [43].

**TABLE 3.** Comparison of voltage and current from OpenDSS and GridLAB-D, for a single power flow solution.

| Maximum error observed in all nodes | Phase |   |   |
|-----------------------------------|-------|---|---|
| Voltage (mV)                       |       | A | B | C |
|                                   | 0.7983 | 0.8102 | 0.7394 |

| Maximum error observed in lines and transformers primary (currents below 0.1 A on OpenDSS are not consider) | Phase |   |   |
|-------------------------------------------------------------------------------------------------|-------|---|---|
|                                                                                                 | A     | B | C |
| Current (mA)                                                                                     | 3.8504 | 0.6626 | 0.9765 |

**V. CONCLUSIONS**

This paper developed synthetic load data derived from real time-varying smart meter data for the Iowa State distribution system test feeder. The smart meter data is from a real distribution system in the U.S. Midwest region. The available smart meter data has an hour resolution, and customers in the same distribution node are aggregated to preserve their privacy. The generated synthetic queueing load data used only the publicly available data that approximate the aggregated behavior of the smart meter data. The generated synthetic load data models individual residential customers and their individual electric assets. The granular appliance-level load is individually known for each of the 1,120 homes for a one-year duration to create the Midwest 240-Node test distribution system. The procedure presented in this paper for the generation of the synthetic load data that aggregated to the complete power system region demand is applicable to other test systems and applicable to other appliance datasets. The procedure consists of analyzing the available demand, addressing possible errors in the load data, segregating the nodal demand to the customer level, and finally generating the loads using the synthetic queueing load model. Assuming portions of the demand are desired to be treated differently, it is only required to remove that demand from the reference given to the queueing load method (e.g., electric vehicles). The studies of this test system with the synthetic load data are intended mainly for smart grid technologies. For this reason, the Iowa State distribution system test feeder OpenDSS model is converted to GridLAB-D and validated in this paper. GridLAB-D is an agent-based approach for simulating smart grids, e.g., market design, building control system design, and integration of new technologies. The GridLAB-D model with the synthetic load data is made publicly available, allowing researchers to validate their methods on a standardized distribution test system.

**APPENDIX**

The appendix analyzes the time-series distribution test system load data (i.e., smart meter data from [35]). The data being analyzed is the first of its type. Having real year long time-series data for an actual distribution feeder. Commonly real data of a test system is not openly available, with the exception of the IEEE European LV test system. However, the data for IEEE European LV is only for a single day. Thus, having data for a complete year is of interest given the privacy and technical challenges in collecting data for customers of an actual system. The year long time-series load data is analyzed and small portions were suspected of being erroneous, and these data are replaced using a generalized linear model. Less than 0.21% of the data was altered. The appendix is separated in three sections. Section -A presents the analyses, evaluating the smart meter data from [35]. Section -B presents the strategy adopted for identifying and addressing the load nodes with strange behavior. A brief discussion on the presented approach is presented in Section -C.

**A. TEST SYSTEM LOAD DATA**

The authors have noticed the presence of significantly small energy consumption for the available data provided in [35]. Fig. 9 presents the number of occurrences of nodal energy consumption of below 100, 10, and 1 Wh. The nodes that have occurrences of below 100 Wh are Nodes 12, 32, 142, 158, 159, and 183. The occurrences of low energy consumption for the expected number of homes are presented in Fig. 10. According to U.S. Energy Information Administration 2015 Residential Energy Consumption Survey [37], homes from the Midwest region have a expected yearly consumption of 9,567 kWh. Assuming the yearly consumption divided by the expected yearly energy consumption the number of res-
idential customers is 1,367, which is considerably different from the 1,120 homes. However, it is also known that the consumption is climate dependent as demonstrated in [33]. Using the month of May for selecting the number of homes in load nodes, the number of homes calculated is 1,161, which is much closer to the 1,120 homes (as described in Section III-E). The estimation of homes in each load node makes use of such an assumption. The energy consumption for the expected homes is computed by dividing the total nodal energy consumption by the expected number of homes. The nodes that have occurrences of below 100 Wh are Nodes 12, 32, 38, 129, 134, 140, 142, 149, 152, 158, 159, 163, 180, and 183. However, the occurrence of such low energy consumption for the node or for the estimated home energy consumption may just be normal operation.

The yearly energy consumption for Nodes 12, 32, 38, 129, 134, 140, 142, 149, 152, 158, 159, 163, 180, and 183 was furthered analyzed. During two periods of time, the year long data has visibly strange behavior. The first is between hours 3,504 to 3,792 at Nodes 158 and 163. The load for these nodes are presented in Fig. 11. The second period is between the hours 6,408 to 6,696 at Nodes 134, 140, 142, 149, 152, 180, and 183. The load for these nodes are presented in Fig. 12.

Figs. 11 and 12 present regions of data that for the nodes of interest do not appear to follow their normal behavior. Fig. 11 shows Node 158 clearly presents two regions of constant energy consumption for over 100 hours. Furthermore, during the same period of time, Node 163 also behaves strangely. In Fig. 12, Nodes 142 and 183 also present two regions of constant energy consumption for over 100 hours. Similarly, during the same period of time Nodes 134, 152, 140, 149, and 180 behave strangely.

The Nodes 158 and 163 are physically close to each other as shown in Fig. 1. Taking a closer look at nearby Nodes 157, 159, 160, 161, 164, and 165, only Node 162 presents strange behavior, it does not appear to be dependent on the test system location.

B. IDENTIFYING NODES AND FILLING IN ERRONEOUS DATA

In this section, a generalized linear model (GLM) is used to fit the expected behavior of a load node, identify nodes that are not behaving as expected, and to fill in the errors of the identified nodes during the erroneous periods. In Section -A, two error periods were identified, i.e., hours 3,504 to 3,792.
FIGURE 13. Energy consumption of Nodes 158, 162, and 163 for the period from 3,504 to 3,792 hours.

and 6,408 to 6,696. Furthermore, some nodes were identified and will be used as the baseline for identification. The only knowledge available for model predictors are the hour of the day and day of the week, i.e., both are classifiers with 24 and 7 possibilities, respectively. The GLM equation considers the interactions of the two predictors as it greatly improves the fitted model. The model provides the average behavior of the load node for the hour of the day and day of the week. Thus, selecting fitting regions for the GLM near the period(s) of interest is expected to provide the average behavior of the error period. The Python package statsmodels [44] was utilized.

A GLM model was fitted for each node for the two periods. The fitting regions are hours 3,144 to 3,480 hour and 3,816 to 4,152 for the first period, and hours 6,048 to 6,384 and 6,720 to 7,056 for the second period. That is, the GLM models were fit using data from two weeks before and two weeks after the erroneous period. The fitted models are first used to compute the expected behavior of the two error periods (i.e., testing regions). The testing regions are from hours 3,600 to 3,700 and 6,450 to 6,550 for the first and second error regions, respectively. The testing regions were chosen empirically from the behavior demonstrated in Section -A. Once the models were fit, the energy percentage error (EPE) and MAPE were computed on the testing region with (3) and (1), respectively, where $t$ is the hour index, $A_t$ is the smart meter load, and $F_t$ is the predicted value from the GLM (i.e., the average hour and day of the week behavior from the fitted GLM model).

The GLM models for the first fitted region presented a minimum, median, and maximum MAPE in p.u. of 0.1197, 0.3167, and 1.5495, respectively. The GLM models for the second fitted region presented a minimum, median, and maximum MAPE in p.u. of 0.0947, 0.2775, and 0.8846 respectively. Please keep in mind that the calculation of MAPE (1) is sensitive to small values of $A_t$, i.e., deviations for small values of $A_t$ have a higher percentage error. Fig. 14 presents the original and GLM model on the fitted region from hours 3,144 to 3,480 and from hours 3,816 to 4,152 for the worst MAPE node of the first fitted region (Node 58), thus illustrating the sensitivity of MAPE to small values of $A_t$.

FIGURE 14. Original and the GLM model energy consumption of the worst load node (Node 58) on the fitted region from hours 3,144 to 3,480 and from hours 3,816 to 4,152.

The performance of the nodes identified were evaluated in Section -A in regard to their EPE and MAPE boundaries to classify problematic nodes. Utilizing the data presented in Table 4 for the performance of the GLM model for the testing region from hours 3,600 to 3,700, possible boundaries using EPE and MAPE were tested and visually analyzed. The lowest EPE and MAPE from Table 4 were utilized as starting points. The resulting classifier for the testing region from hours 3,600 to 3,700 became a combination of presenting nodes with EPE larger than 0.67 and MAPE larger than 1.75, resulting in adding Nodes 41 and 154, which were verified visually, as presented in Fig. 15.

Similarly, utilizing the data presented in Table 5, the performance of the GLM model for the testing region from hours 6,450 to 6,550 for EPE and MAPE were evaluated. Utilizing the experience from the first testing region and visually testing multiple boundaries the same classifier was made. Nodes with identified problems are classified by the combination of presenting EPE larger than 1.99 and MAPE larger than 4.47. No nodes were been added.

The selection of nodes identified as errors were performed
TABLE 4. Testing region from hours 3,600 to 3,700 EPE and MAPE for Section -A identified nodes and added classified nodes.

| Node | EPE (p.u.) | MAPE (p.u.) |
|------|------------|-------------|
| 158  | 1.0446     | inf         |
| 162  | 0.8734     | 1.9536      |
| 163  | 1.2984     | 5.0156      |

| Added node | EPE (p.u.) | MAPE (p.u.) |
|------------|------------|-------------|
| 41         | 1.8870     | 2.9512      |
| 154        | 2.4025     | 2.9391      |

TABLE 5. Testing region from hours 6,450 to 6,550 EPE and MAPE for nodes identified in Section -A.

| Node | EPE (p.u.) | MAPE (p.u.) |
|------|------------|-------------|
| 134  | 2.2388     | 5.0973      |
| 140  | 4.1193     | 1.0915 x 10^1 |
| 142  | 7.1508     | 6.1649 x 10^21 |
| 149  | 2.3612     | 5.0088      |
| 152  | 2.1967     | 4.6794      |
| 180  | 2.6942     | 5.1156      |
| 183  | 10.3285    | 4.4456 x 10^22 |

FIGURE 15. Energy consumption of the Nodes 41, 154, 158, 162, and 163 for the period from 3,504 to 3,792 hours.

FIGURE 16. Model energy consumption of Nodes 41, 154, 158, 162, and 163 for the period from hours 3,504 to 3,792.

FIGURE 17. Model energy consumption of Nodes 134, 140, 142, 149, 152, 180, and 183 for the period from hours 6,408 to 6,696.

using the GLM model in regards to their EPE and MAPE. The replacing of the data suspected of being erroneous is performed by the same GLM model. The first is between hours 3,504 to 3,792 at Nodes 41, 154, 158, 162, and 163. The new load created for these nodes is presented in Fig. 16. The second period is between hours 6,408 to 6,696 at Nodes 134, 140, 142, 149, 152, 180, and 183. The new load for these nodes is presented in Fig. 17. Comparing Fig. 16 with Fig. 15 and Fig. 17 with Fig. 12, the differences of the original data with the replaced model data are presented.

C. DISCUSSION

Collecting smart meter data for 1,120 customers for a year long period is not trivial, e.g., data collection equipment is subject to environmental hardships, possibility of equipment failure, communication failure, and misuse of equipment. Given that the environment of data collection is a distribution system in the Midwest U.S. (i.e., Iowa) many individuals had access to the collection equipment (i.e., not a controlled environment). The analyzed time-series distribution test system load data presented shows two time periods where some of the load nodes presented strange behavior. The first time period was from hours 3,504 to 3,792 for Nodes 41, 154, 158, 162, and 163. The second time period was from hours 6,408 to 6,696 for Nodes 134, 140, 142, 149, 152, 180, and 183. The strange behavior was not limited to nodes presented here, however the strange behavior only suggested possible errors with the data. The authors attempted to preserve as much of the original data as possible, avoiding replacing correct data from the two time periods of strange behavior. The presented method of replacing data used GLM models on the selected nodes and regions, and altered less than 0.21% of the total...
REFERENCES

[1] Fatih Birol, "Digitalization & Energy," International Energy Agency (IEA), Tech. Rep., 2017.

[2] N. Good, K. A. Ellis, and P. Mancarella, “Review and classification of barriers and enablers of demand response in the smart grid,” Renewable and Sustainable Energy Reviews, vol. 72, pp. 57 – 72, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032117300436

[3] A. B. Birchfield, E. Schweitzer, M. H. Athari, T. Xu, T. J. Overbye, A. Scaglione, and Z. Wang, “A metric-based validation process to assess the realism of synthetic power grids,” Energies, vol. 10, no. 8, p. 1233-2017.

[4] 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 Transactions on Power Systems, vol. 32, no. 4, pp. 3528-3625, 2017.

[5] B. P. Bhattarai, B. Bak-Jensen, P. Mahat, and J. R. Pillai, “Voltage controlled dynamic demand response,” in IEEE PES ISGT Europe 2013, IEEE, 2013, pp. 1-5.

[6] T. M. Hansen, E. K. P. Chong, S. Suryanarayanan, A. A. Maciejewski, and A. J. Ladd, “Enabling grid scale demand response through a novel transactive energy systems framework,” in Digitalization, Energy, and Operations in the Smart Grid, 2019 IEEE Power & Energy Society General Meeting (PESGM), 2019.

[7] L. Park, Y. Jung, S. Cho, and J. Kim, “Residential demand response for renewable energy resources in smart grid systems,” IEEE Transactions on Industrial Informatics, vol. 13, no. 6, pp. 2864–2875, 2017.

[8] C. Mateo, G. Prettico, T. Gómez, R. Cossent, F. Gangale, and F. Postigo, “A metric-based validation process to assess the realism of synthetic power grids,” Energies, vol. 10, no. 8, p. 1233-2017.

[9] Z. Wang, “A multi-timescale control of demand flexibility in smart distribution networks,” Energies, vol. 10, no. 1, p. 37, 2017.

[10] B. P. Bhattarai, K. S. Myers, and J. W. Bush, “Reducing demand charges and onsite generation variability using behind-the-meter energy storage,” in 2016 IEEE Conference on Technologies for Sustainability (SusTech), IEEE, 2016, pp. 140–146.

[11] B. P. Bhattarai, K. S. Myers, B. Bak-Jensen, and S. Paudyal, “Multi-timescale control of demand flexibility in smart distribution networks,” Energies, vol. 10, no. 1, p. 37, 2017.

[12] N. Good, K. A. Ellis, and P. Mancarella, “Review and classification of barriers and enablers of demand response in the smart grid,” Renewable and Sustainable Energy Reviews, vol. 72, pp. 57 – 72, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032117300436

[13] A. B. Birchfield, E. Schweitzer, M. H. Athari, T. Xu, T. J. Overbye, A. Scaglione, and Z. Wang, “A metric-based validation process to assess the realism of synthetic power grids,” Energies, vol. 10, no. 8, p. 1233-2017.

[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 Transactions on Power Systems, vol. 32, no. 4, pp. 3528-3625, 2017.

[15] B. P. Bhattarai, B. Bak-Jensen, P. Mahat, and J. R. Pillai, “Voltage controlled dynamic demand response,” in IEEE PES ISGT Europe 2013, IEEE, 2013, pp. 1-5.

[16] T. M. Hansen, E. K. P. Chong, S. Suryanarayanan, A. A. Maciejewski, and A. J. Ladd, “Enabling grid scale demand response through a novel transactive energy systems framework,” in Digitalization, Energy, and Operations in the Smart Grid, 2019 IEEE Power & Energy Society General Meeting (PESGM), 2019.

[17] L. Park, Y. Jung, S. Cho, and J. Kim, “Residential demand response for renewable energy resources in smart grid systems,” IEEE Transactions on Industrial Informatics, vol. 13, no. 6, pp. 2864–2875, 2017.

[18] C. Mateo, G. Prettico, T. Gómez, R. Cossent, F. Gangale, and F. Postigo, “A metric-based validation process to assess the realism of synthetic power grids,” Energies, vol. 10, no. 8, p. 1233-2017.

[19] Z. Wang, “A multi-timescale control of demand flexibility in smart distribution networks,” Energies, vol. 10, no. 1, p. 37, 2017.

[20] B. P. Bhattarai, K. S. Myers, and J. W. Bush, “Reducing demand charges and onsite generation variability using behind-the-meter energy storage,” in 2016 IEEE Conference on Technologies for Sustainability (SusTech), IEEE, 2016, pp. 140–146.

[21] B. P. Bhattarai, K. S. Myers, B. Bak-Jensen, and S. Paudyal, “Multi-timescale control of demand flexibility in smart distribution networks,” Energies, vol. 10, no. 1, p. 37, 2017.
[41] W. H. Kersting, *Distribution system modeling and analysis*. CRC press, 2012.
[42] “GridLAB-D model with the synthetic load data is made publicly available,” will be published on paper acceptance. The authors are selecting the venue.
[43] “GridLAB-D model with the synthetic load data is made publicly available,” 2020. [Online]. Available: db.bettergrids.org
[44] S. Seabold and J. Perktold, “statsmodels: Econometric and statistical modeling with Python,” in 9th Python in Science Conference, 2010.

**FERNANDO BERETA DOS REIS** (GS’17–M’21) received his B.A.Sc. degree in Electrical Engineering, in 2014 from PUCRS (Pontifícia Universidade Católica do Rio Grande do Sul), Brazil; his M.Sc. in Electrical Engineering in 2016 from UFSC (Universidade Federal de Santa Catarina), Brazil; and, his Ph.D. in Electrical Engineering in 2020 from South Dakota State University, Brookings, USA. He is currently a Research Scientist/Engineer at the Pacific Northwest National Laboratory, USA. His research interests include power systems modeling; grid integration of sustainable energy technologies; transactive energy; and power systems simulations and co-simulations.

**REINALDO TONKOSKI** (S’04–M’11–SM’18) received his B.A.Sc. degree in Control and Automation Engineering, in 2004 and his M.Sc. in Electrical Engineering in 2006 from PUC-RS (Pontifícia Universidade Católica do Rio Grande do Sul), Brazil; and, his Ph.D. in 2011 from Concordia University, Canada. He was with CummeR-ERGY, Natural Resources Canada, from 2009 to 2010 where he worked on projects related to the grid integration of renewable energy sources. He is currently an Associate Professor in the Electrical Engineering and Computer Science Department, South Dakota State University, Brookings, USA and a Visiting Faculty at Sandia National Laboratories, Albuquerque, NM. Dr. Tonkoski has authored over ninety technical publications in peer reviewed journals and conferences and is currently an Editor of IEEE Transactions on Sustainable Energy, and Associate Editor of IEEE Access and IEEE Systems Journal. His research interests include grid integration of renewable energy systems and batteries, smart grid, power electronics and controls.

**BISHNU P. BHATTARAI** (S’12, M’15, SM’18) received a Ph.D. degree in Electrical Engineering from Aalborg University, Denmark, in 2015. He is currently a Senior Research Scientist/Engineer at the Pacific Northwest National Laboratory, USA. His research interests include advanced modeling and simulation of power distribution systems; transactive energy system; and power and communication co-simulations. He was the recipient of the Gold Medal from the ‘President of Nepal’ in 2011; ‘Green Talent Awards’ from the ‘German Federal Ministry of Education and Research’ in 2017; and best Reviewer Awards from “Elsevier IJEPES” in 2017 and from “IEEE Transactions on Smart Grid” in 2018 and 2019. He is currently serving as an Editor for IEEE Transactions on Smart Grid.

**TIMOTHY M. HANSEN** (S’07–GS’11–M’15–SM’20) received the B.S. degree in computer engineering with high honors from the Milwaukee School of Engineering, Milwaukee, WI, USA, in 2011, and the Ph.D. degree in electrical engineering from Colorado State University, Fort Collins, CO, USA, in 2015.

In 2014–2015, he held a graduate research position at the Distributed Energy Systems Integration group at the National Renewable Energy Laboratory, Golden, CO, USA. He is currently an Assistant Professor with the Electrical Engineering and Computer Science Department, South Dakota State University, Brookings, SD, USA. His research interests are in the application of optimization, high-performance computing, and electricity markets to the areas of sustainable power and energy systems, low-inertia power systems, smart cities, and cyber-physical-social systems.

Dr. Hansen is currently the IEEE Region 4 Siouxland Section Chair and an active member in the IEEE Power and Energy Society Power Engineering Education Committee. He is a member of ACM SIGHPC and SIGAPP, and is currently an Associate Editor of Elsevier Sustainable Computing: Informatics and Systems.