California Test System (CATS): A Geographically Accurate Test System based on the California Grid

Sofia Taylor*, *Student Member, IEEE,* Aditya Rangarajan*, *Student Member, IEEE,* Noah Rhodes, *Student Member, IEEE,* Jonathan Snodgrass, *Member, IEEE,* Bernie Lesieutre, *Fellow, IEEE* Line A. Roald, *Member, IEEE*

Abstract—This paper presents the California Test System (CATS), a synthetic transmission grid in California that can be used by the public for power systems research without revealing any critical energy information. The proposed synthetic grid combines publicly available geographic data of California's electric infrastructure, such as the actual location of transmission corridors, with invented topology and transmission line parameters that are "realistic but not real". The result is a test system that is suitable for power flow and optimal power flow analysis. The methods used to develop and evaluate the CATS grid are documented in detail in this report.

Index Terms—test system, synthetic grid, geographic information systems, optimal power flow.

I. INTRODUCTION

Easy access to realistic data sets is widely recognized as a catalyst for impactful research. Realistic power systems data enable the development and benchmarking of effective algorithms for system operations and planning, allowing researchers to move beyond small-scale academic examples to assess scalability and benefits of new methods on more realistic systems. In addition, there is growing recognition that we must prepare the grid for more frequent extreme weather events, including increased wildfire risk, more destructive hurricanes, tornados, and flooding, and heat waves that may elevate electric load. Modeling and understanding these impacts requires us to consider not only the electric grid, but also the environmental and social context around it. Developing new tools for operations and planning that capture this context requires access to geographically accurate grid data that can be correlated with other public data sources, such as data on wildfire and flooding risks, and information on the vulnerability of the population to power outages and weather impacts. This paper presents a new, large-scale test case that addresses the need for geographically accurate and electrically realistic grid models.

There are important reasons to keep the true electric grid model from public view. For example, details about electricity generation, transmission, and distribution might compromise the competitiveness and privacy of markets. Additionally, such knowledge could facilitate cyber and physical attacks on electric infrastructure, which pose a threat to public health and national security. To fortify the privacy and security of electric infrastructure in the United States, the Federal Energy Regulatory Commission (FERC) has classified certain power system information as Critical Energy/Electric Infrastructure Information (CEII) [1]. As a result, CEII data and models of power systems are not publicly available. They can be accessed by researchers that obtain permission, but can not be published along with test case results, creating challenges for benchmarking and sharing research. Security concerns aside, real grid models tend to be very complex and often have missing data, making them challenging to use.
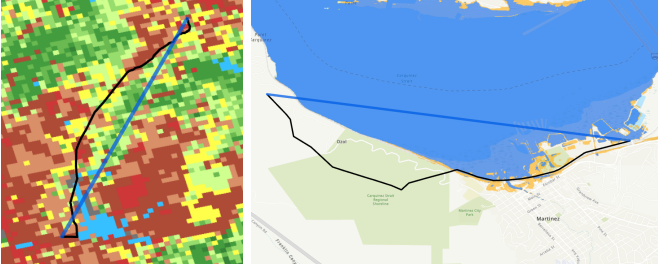
Thus, to fullfill research needs, researchers have developed synthetic grid models with realistic, but synthetic parameters. Synthetic grids include a set of electrical components, including transmission lines, transformers, buses, substations, generators, and loads, with defined parameters and connections that determine how current will flow through the system. The classic IEEE test cases were among the first synthetic grid models [2], followed by recent synthetic networks created by Texas A&M University [3] and University of Wisconsin-Madison [4], the Reliability Test System-Grid Modernization Lab Consortium (RTS-GMLC) network [5], and other test cases for benchmarking AC optimal power flow algorithms [6]. These synthetic grids are typically easy to use relative to real grid models and allow users to freely share results.

One shortcoming of current synthetic grids is their lack of realistic geographic information. Though inspired by real electric grids, the IEEE test cases do not include any geographic information. In contrast, the recently developed grids [3]–[5] all contain geographic coordinates that span various regions across the United States. These models do not represent any particular real power system, but model the characteristics and electrical behavior of real power systems [5], [7]–[9]. Thus, the transmission lines, substations and buses do not correspond to any existing power equipment. Also, the transmission lines are represented as straight lines connecting two nodes, rather than nonlinear paths that curve based on local topography, vegetation, and property ownership.

Realistic or accurate geography with high granularity is important for geo-referenced applications, including studies related to weather, topography, and socio-economic considerations. For example, the models in [10] balance the competing

(a) A 230kV line in northern California overlaid on a wildfire risk map.

(b) A 60kV line in the Bay Area overlaid on NOAA 5-foot sea level rise projections.

Fig. 1: Images depicting the usefulness of the accurate paths of power lines (in black) versus their point-to-point approximations (in blue) in analyses with geo-referenced data.

risks associated with wildfire ignition from power equipment and preemptive power shutoffs. Using a geographically accurate synthetic grid to test such a model allows for synthesis with spatial data, including wildfire risk maps and demographic data. Thus, there is a need for synthetic grids without CEII-protected information that represent the existing geography of real power systems. This paper aims to address this need.

The main contribution of this paper is the California Test System (CATS), a geographically accurate synthetic transmission network model that is geo-located in the state of California. The starting point for this model was the true geographic data of California's electric infrastructure, which is publicly available through the California Energy Commission (CEC) [11] and the Energy Information Agency (EIA) [12]. However, while there is an abundance of publicly available data, there several important parts of the data that are missing. This includes the topology of the system (e.g., connections between components), line and generation parameters, and locations and parameters of transformers and reactive power compensation. Thus, we supplemented the available geographical grid information with the necessary synthetic data to create a geo-located test system suitable for power flow and optimal power flow studies. To achieve this, we added approximate substation and node topologies to the system, leveraged additional generation and load data from the EIA [12] and California Independent System Operator (CAISO) [13], and assigned realistic line parameters based on publicly available data from FERC [14]. The result is an open-source, non-CEII transmission network model suitable for geo-located applications and large scale power flow and optimal power flow studies. The CATS grid model is available in a GitHub repository in MATPOWER and GIS formats [15].

In summary, the contributions of the paper are twofold. First, we provide a new, openly available test system for (optimal) power flow studies that allow for large-scale analysis and correlation with other geo-referenced data. The test system has been evaluated for a full year of load and generation data, and is made easily accessible to researchers and the general public through our GitHub repository [15]. Second,

we describe the procedure that we used to create the CATS grid model. While the description primarily serves to explain how this specific grid model was developed, the procedure can be adapted to create similar transmission and distribution models in other geographic locations.

The rest of this paper is organized as follows. Section II details the data sources and Sections III-VI describe the methodology used to create the test system. Section VII presents evaluation metrics and performance results. Sections VIII and IX summarize the contributions, limitations, and future work.

II. DATA SOURCES

The CATS synthetic grid was developed using several sources of publicly accessible data, which are organized in the following categories: electric infrastructure geographic information, generation data, load profiles, and transmission line parameter data.

A. California Electric Infrastructure Geographic Data

Geographic information systems (GIS) use maps to represent spatial data. We obtained GIS data of California's power lines and substations, including locations and attributes, from the CEC [11].

While the CEC provides GIS data of California's power plants, the EIA also publishes information about California's power plants, including geographic coordinates, in the Form EIA-860 [12]. The location and capacity of the generators and plants from the CEC and EIA sources are similar, but not exactly the same. To inform generator locations in the CATS network, we selected the 2019 Form EIA-860 over the CEC power plant data since it provides additional attribute fields, such as power factor, that could be useful for various applications.

B. Generation Data

In addition to geographic coordinates, the 2019 Form EIA-860 contains useful generator attribute information, such as the nameplate MW capacity, power factor, minimum MW values, fuel type and unit type. However, it does not contain details about renewable energy generation output and non-renewable generation cost curves. Since we need such information to create the synthetic grid, we use 2019 state-wide renewable data published by CAISO [13] and quadratic cost curve coefficients from [16]. The methods used by [16] are as follows. To create the cost curves for steam boiler plants, [16] used typical heat rate values from [17] and multiplied them by fuel costs from the EIA website. For natural gas plants, heat rates from GE are multiplied by average natural gas fuel prices. For nuclear, wind and solar, offer curve data were used. The generators in the Form EIA-860 were grouped into clusters, and representative generators from each cluster were manually assigned cost curves. The remaining generators were automatically assigned cost curves based on the representative generators [16].

C. Load Data

To generate load profiles, we leverage publicly available aggregate hourly load data from CAISO for 2019 [13] in combination with a method developed as part of the EPIGRIDS project, which disaggregates state-wide temporal load profiles to individual loads. The output of this method is hourly load data at census tract level granularity which captures geographic variation throughout California [18], [19]. In the EPIGRIDS method, GIS data is used to estimate the percentage of the state load consumed at each load location, as well as the distribution of residential, commercial, and industrial load types. This is combined with information on typical load profiles associated with the three load types to create a load profile for each load bus, appropriately scaled to match the total state-wide load.

D. Transmission Line and Transformer Parameter Data

Transmission line parameters, including resistance, reactance and susceptance, are protected data and not available in public data sets. Since these parameters are needed for grid analysis, we use publicly available data to generate realistic (but synthetic) parameters for the transmission lines and transformers in the CATS network. FERC publishes historical and current annual reports of information about electric utilities in the United States. Two of these reports, the Form No. 1 "Annual Report for Major Electric Utility" [14], and the Form No. 715 "Annual Transmission Planning and Evaluation Report" [20], contain useful data for assigning line parameters. The Form No. 1 is publicly available, while the Form No. 715 is classified as CEII and no longer released to the public. Although access to the Form No. 715 can be requested, the authors opted instead to only use the Form No. 1 from 2010 and previously published average values and statistical data derived from the Form No. 715 [8], [19] to avoid concerns regarding protected data in the CATS grid model.

The Form No. 1 was created for utilities to report their bulk electric system assets to FERC for annual accounting. In California, three utility companies submitted the Form No. 1 report and thus their data is utilized to create the proposed grid. The data includes voltage level, transmission line length, number of conductors per phase, conductor size, conductor type (ACSR, ACSS, XLPE, etc.), transmission structure material, and construction type. The granularity of the data varies, ranging from total mileage of transmission lines at each voltage level to individual transmission structure construction for individual segments of each transmission line.

The Form No. 1 does not contain any useful data for transformer parameters. We therefore leverage average per unit impedance values using the transformer base MVA for each pair of primary-secondary voltages from [19]. The X/R ratios for all transformers with MVA values ranging from 50 MVA to 2000 MVA were obtained from [8].

III. TOPOLOGY

While the data from CEC and EIA provide accurate information about the geographical location of individual components, they do not provide a description of how the components are connected, i.e., they lack important information regarding the system topology. To create a fully connected network suitable for power system simulation and analysis, we introduce a method for connecting components and describe various data cleaning and processing steps.

A. Initial topology

The available datasets include substation locations, generator locations, and transmission line paths. The first step in achieving a connected topology is to assign substation connections to the end points of the transmission lines. To do this, we calculate the distance from a transmission line endpoint to each substation in the network, assign the closest substation as a connection, and repeat for each transmission line endpoint in the network. Similarly, for assigning generators to substations, we calculate the distance from each generator to each substation and assign a generator connection to the closest substation.

While this connectivity method is simple and intuitive, several data challenges cause the resulting network to be a poor representation of the CAISO network. Theses challenges and their solutions are discussed below.

B. Connecting transmission lines to substations

The transmission lines from the CEC data represent the actual line locations, but the data does not accurately model electrical circuits and connections. For example, many transmission lines are represented as several line segments. These line segments should not be connected to substations, but rather to each other. If we force all line ends to connect to a substation, many of the segments are connected incorrectly.

To remedy this challenge, additional nodes were required to represent electrical interconnections between transmission line segments. These "added nodes" are placed at every transmission line segment endpoint that does not already have a node or substation within a small search radius. A conservative search radius of just 12 meters ensures that nodes are still placed at the endpoints of very short line segments that exist within cities. The result of this procedure is thousands of added nodes, but a much more accurate topology.

C. Transmission lines that branch into separate paths

In the GIS data, it was common for transmission lines to have other lines branching away along their length. An example is shown in Fig. 2. Per the method described above, a new node would be added at the branching line's endpoint. However, the main line often would not terminate at the node, and therefore will not be modeled as an electrical connection. In this case, the main line must be segmented into two parts to correctly model the connection to the branch. Thus, we segment all such lines into smaller line paths to create the appropriate end points and connections between the circuits. Unfortunately, the process created a few hundred extremely short line segments, some smaller than one meter. These lines segment were manually deleted or merged with another line.

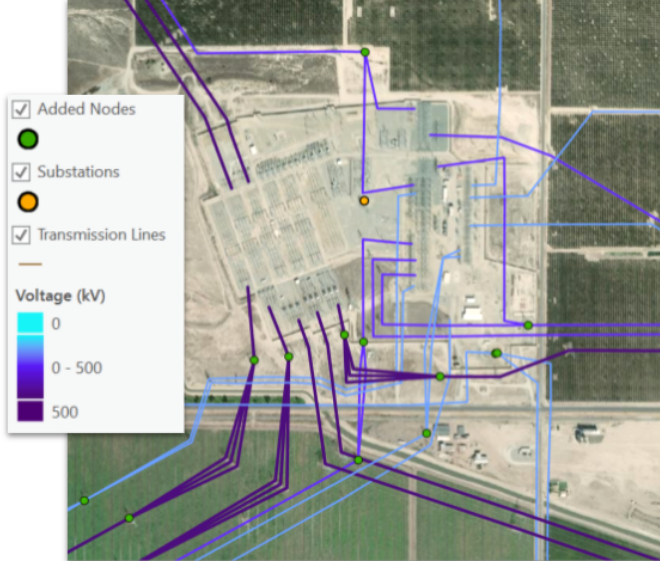Fig. 2: Transmission line branching off a from another line.

Fig. 3: GIS representation of Midway Substation in Kilowatt, California, overlaid on top of a Google Maps satellite image.

D. Manually resolving topology issues

After the above steps are complete, the network connectivity process (where all lines are connected to the closest node) can be run again. However, although the above methods resolve a lot of issues, they also cause secondary topology problems that require manual analysis and data cleaning. Certain parts of the topology, such as Midway Substation in Kilowatt, California, shown in Fig. 3, are not correctly connected because too many nodes have been added to nearby transmission line endpoints (due to the conservative search radius used when adding new nodes). Large substations like Midway cover a large area of land in reality but are represented as a single point, resulting in relatively large distances between the substation point and adjacent line endpoints. This led to the addition of many extra nodes, and many lines connect to nodes outside of the substation instead of to the substation itself. To ensure that the lines were connected correctly to the substations, these extra nodes at the Midway substation were removed manually.

To aid in identifying other locations where there may be connectivity challenges, we created a tool to calculate the graph-distance (with line length as edge weights) and geographic-distance between node pairs. A large discrepancy between these values could indicate that nearby nodes are not electrically connected when they should be. This allowed easier identification of locations similar to Midway substation, where creating correct topology connections in a fully automated way is challenging. After using the tool and inspecting

to identify a connectivity issue, we manually modified the geographic data to solve the problem.

One particular connectivity challenge arises when a line is assigned a connection to the same node at both ends. In these cases, an added node at an interconnection was often closer than the substation node, and therefore both line endpoints were assigned to that same added node. We produced a list of these lines and manually addressed them by deleting nodes and merging, extending, and deleting lines in ways that allowed components to correctly connect.

E. Substation transformers

The final topology challenge is the addition of transformers. The datasets include line voltage levels, but not connections between voltage levels. We must create substation topologies if multiple voltage levels connect to the same substation. Thus, when multiple line voltage levels connect at the same node, we split the node into multiple nodes with transformers in between them and then reassign lines and generators.

The process for transformer node creation is as follows.

1) Identify the number of voltage levels V at a node.
2) Add $V - 1$ additional nodes to the network.
3) Add new transformer branches to connect the nodes, where the highest voltage node connects to the second highest voltage node, the second highest voltage node connections to the third-highest, etc.
4) Assign each line that connected to the original node to the new node representing the correct voltage level.
5) Assign generators to the highest voltage level in the set.

F. Final topology

For each of the topology creation steps, we modified and cleaned some of the input data and then re-ran the automated connectivity steps. The final grid topology is shown in Fig. 4.

IV. Assigning generator, load and renewable energy data

Next, we describe our methodology for assigning generation, load and renewable energy data in the CATS.

1) Assigning data for conventional generators: The Form EIA-860 files include static attributes, including the nameplate capacity, nameplate power factor, latitude, and longitude.

In addition to the generator locations and capacities, we also need cost curves to model generator dispatch within the electricity market. The cost curve coefficients (see Section II-B) are correlated with the 2019 Form EIA-860 generators. For 'Plant Codes' and 'Generator ID's' that directly matched between the 2010 cost curve spreadsheet and the 2019 generators, the cost curve coefficients were simply copied over. For the remaining non-renewable generators, we assign the coefficients of the closest-sized generator of the same type. For the renewable generators, we assign coefficients of zero.
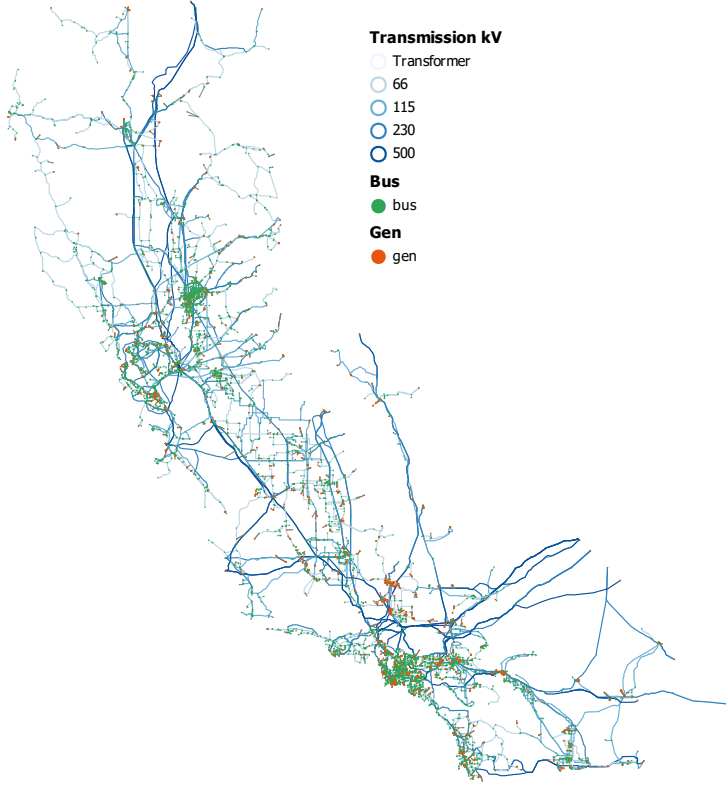
Fig. 4: California synthetic grid topology.

Fig. 5: Flowchart of the grid creation process.

2) Assigning data for renewable generation: To account for the variability of solar and wind generation, we used the state-wide generation data published by CAISO for 2019 [13]. The data consists of load and generation from different resources for every five minute interval, from which we sample points at the beginning of every hour. Before beginning, we scaled the capacities of all the solar generators to account for the lower solar capacity in the grid topology created, compared to the peak solar generation in California. The next step is to assign a portion of the total state-wide generation from renewable sources to each renewable generator in the CATS. The capacity of each generation was scaled according to its actual rating, as shown in equations (1a) and (1b).

$$P_{PV,i}^{max,s} = \frac{P_{PV,i}^{max}}{\sum_{j \in \mathcal{PV}} P_{PV,j}^{max}} P_{PV}^{s} \quad \forall i \in \mathcal{PV}, s \in \mathcal{S} \quad (1a)$$

$$P_{W,i}^{max,s} = \frac{P_{W,i}^{max}}{\sum_{j \in \mathcal{W}} P_{W,j}^{max}} P_{W}^{s} \quad \forall i \in \mathcal{W}, s \in \mathcal{S} \quad (1b)$$

In the above equations, $\mathcal{PV}$ is the set of all solar generators and $\mathcal{W}$ is the set of all wind generators in the CATS. The set $\mathcal{S}$ denotes the generation scenarios considered. $P_{g,i}^{max}$ is the actual capacity of generator i (scaled to ensure that the grid
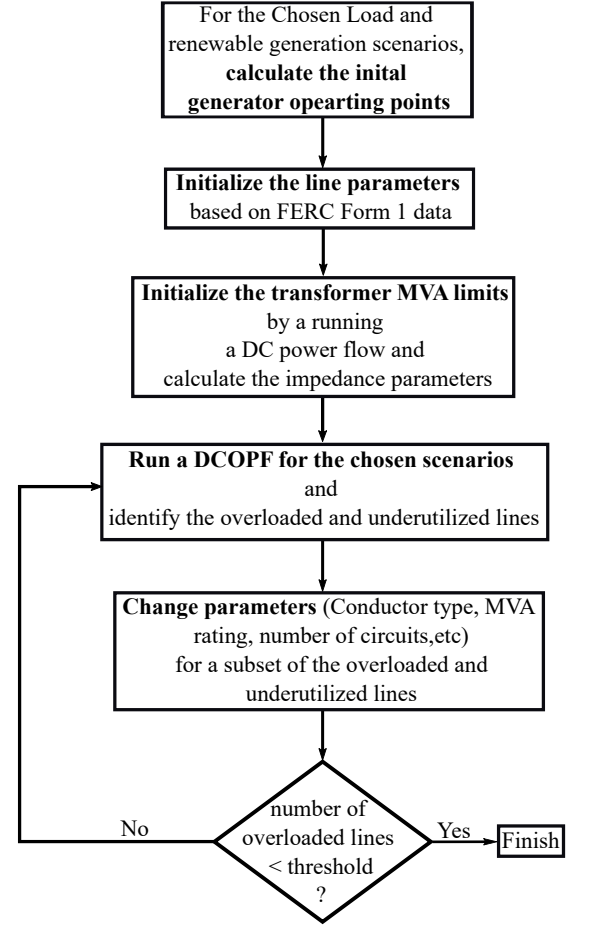
is able to produce peak solar), P_g^s represents the total amount of solar or wind generation in the system, respectively, and $P_{g,i}^{max,s}$ is scaled capacity proportional to the generation from renewable sources in scenario s.

We note that the above assignment policy keeps the ratio of renewable generation constant between generators over time. In future work, we could improve the renewable energy production scenarios to account for more geographic variability.

3) Load: We utilize hourly load data scenarios from the EPIGRIDS project [18], [19]. These loads are given per census tract, and therefore need to be assigned to specific locations in the grid. A first approximation is to assign each load to the nearest CEC substation. However, this method would leave many substations with very high load, many with no load, and some radial substations with no load or generation. For more realistic load assignment, we use an optimization assignment problem, shown in Problem 2, that minimizes the distance between the EPIGRIDS load buses and the CEC substations.

$$\min_{x} \quad \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{N}} x_{ij} c_{ij} \quad (2a)$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{N}} x_{ij} = 1 \quad \forall i \in \mathcal{L} \quad (2b)$$

$$\sum_{i \in \mathcal{L}} x_{ij} \geq 1 \quad \forall j \in \mathcal{N} \quad (2c)$$

Here, the load assignment variable x_{ij} is a binary variable that determines if load i is assigned to node j. The objective function minimizes the cumulative distance from the location of each load to the substations, represented by c_{ij}. The set $\mathcal{L}$ is the set of all loads from EPIGRIDS, with one load per census tract. The set $\mathcal{N}$ is the set of nodes where we require the algorithm to assign at least one load. This set $\mathcal{N}$ includes the original substations from the CEC dataset, as well as the added nodes located at the end of radial branches that do not have generators attached. The constraints on the load assignment are that each load $i \in \mathcal{L}$ is assigned is assigned to exactly one node (2b), and that each node $j \in \mathcal{N}$ is assigned at least one load (2c). The set of loads $\mathcal{L}$ is larger than the set of nodes $\mathcal{N}$ at which we assigned the loads, permitting a feasible solution to the problem. Since this problem satisfies the conditions for total unimodularity, it produces a solution with $x_{ij} \in \{0, 1\}$ without explicitly enforcing that x_{ij} is binary.

V. Assigning Line and Transformer Parameters

To obtain a system which gives rise to feasible and realistic power flow (PF) and optimal power flow (OPF) solutions, we need to add realistic line and transformer parameters, such as impedance and MVA limits. This section describes the procedure used for generating these limits. Key steps are outlined in Fig. 5.

A. Input data

1) Choosing load and generation scenarios: To create a realistic power system network model that is feasible for a wide range of load and renewable generation scenarios, we have to consider more than one loading condition when assigning line parameters. Considering every possible load and generation scenario is not practical for computational reasons. To increase the computational efficiency of the grid creation algorithm while still designing for a large range of operating conditions, the load data was sub-sampled to create a representative set of scenarios that approximates the boundaries of the set of load and generation scenarios.

Specifically, we picked a total of 245 loading scenarios. The first 121 scenarios are chosen to be the hour with maximum load, as well as the 60 hours before and after. The next 121 scenarios are chosen to be the hour with the minimum load, as well as the 60 hours before and after. In addition, we consider 3 scenarios that represent the hour with the maximum solar generation, the hour with maximum wind generation and the hour with the overall lowest renewable generation. This ensures that we observe a range of both load and renewable generation scenarios, as well as different hours of the day and days of the week.

2) Generate initial generation profiles for each scenario: When creating generation profiles for the scenarios, we want to reflect typical operating conditions (i.e., conditions that allow the lowest cost generators to run). At the same time, we should avoid over-optimizing the grid such that transmissions lines are sized only to support the lowest cost generator dispatch (economic dispatch). Also, electricity demand changes throughout a day or week, as well as across several months or years. If

the grid creation process does not exhibit enough variety in the generator dispatch and unit commitment, the grid will not be able to support varying load flow patterns.

Past experience has also shown that synthetic grid models are highly sensitive to the set of generators committed, and that each generator must be dispatched at it's maximum output in at least one scenario [19]. Otherwise, the transmission lines connected to the generator points of interconnection (POI) for decommitted generators will be inadequately designed.

To obtain a varied set of power injections from the generators, we generate two sets of generation schedules for each hour using two methods we call the *economic dispatch* and *uneconomic dispatch*:

- *Economic dispatch:* For each hour, we implement a simple unit commitment algorithm. This algorithm iteratively decommits the most expensive generator until a target spinning reserve level of 10% is met. Once the unit commitment is fixed, the generators are dispatched using an economic dispatch algorithm that minimizes the generator cost subject to the total demand equaling total generation.
- *Uneconomic dispatch:* In the uneconomic dispatch, we follow a similar procedure, but instead decommit the cheapest generators to create an "uneconomic dispatch". The final injections are again computed by running an economic dispatch algorithm considering just the most expensive generators.

The economic and "uneconomic" dispatch scenarios result in all generators being dispatched at, or near, their maximum power output in at least one generator unit commitment scenario. Since we create two generation scenarios for each of our 245 load scenarios, we consider a total number of 490 power injection scenarios.

3) Initializing the transmission line and transformer parameters: As a final input to our method, initial transmission line and transformer impedances are assigned to each of the network branches. Instead of using a simple assignment such as assigning a uniform per-unit-length impedance to all transmission lines, we used transmission line data from the FERC Form 1 to make the initial assignment. For each transmission line at each voltage level in the CATS, the line in the FERC Form 1 with the closest length was identified. If the utility company listed in the CEC data matched the utility company listed in the FERC Form 1, only the lines in the Form 1 data corresponding to that utility company were examined.

We used the Form 1 data for the matched CEC transmission lines, including conductor size (in kcmil), conductor type, and number of conductors per phase, to determine ampacity limits and transmission line impedances for the corresponding transmission line in the CATS, following the methodology described in [19]. As part of this process, transmission line manufacturer's data sheets [21] are used to determine ampacity limits, while [22] is used to determine approximate GMR and GMD values which are then used to compute synthetic per-unit length transmission line impedances for the lines in the Form 1. Additionally, the calculated ampacity limits are multiplied by the rated voltage of the transmission line to

calculate probable MVA thermal ratings for each transmission line in the Form 1.

By examining geographic regions such as states or approximate ISO or RTO service territory, we can determine MVA ranges for each voltage level and region. This data is validated using the MVA limit ranges for transmission lines at each voltage level in [8], [19]. We combine the calculated per-unit-length impedance parameters with the MVA limit ranges to produce a table of possible conductors configurations for each transmission line. Later, we use this table to adapt the transmission line parameters as the MVA limits of lines are increased or decreased, as explained below.

The MVA values and ranges were created assuming that all conductors are ACSR [21]. However, if the conductor material type is changed to ACSS, the ampacity roughly doubles, without a substantial change to the GMR of the wire [23]. This allows the MVA rating of a transmission line to increase by up to 100% of the original value without modifying the corresponding R,X,B values.

Since the Form 1 does not contain useful transformer data, we assign an initial limit of 2000 MVA to each transmission-level transformer. The average per unit impedance values are obtained from [19] using the transformer base MVA and the primary-secondary voltages, and corresponding X/R ratios were obtained from [8]. Once the initial line and transformer parameters are calculated, we solve a DC power flow for all 490 scenarios to calculate the resulting flows through all transmission lines and transformers. We then resize the transformers to have an MVA limit equal to the maximum value calculated from the DC power flow and recompute the impedance parameters corresponding to these calculated flows.

B. Algorithm for Updating Transmission Line Parameters

The initial line parameter assignment is typically inaccurate and the generation schedules obtained with the uneconomic and economic dispatches, which do not account for network constraints, may not be feasible. In the following section, we describe our algorithm for adjusting both the generation dispatch and the line parameters to obtain a system that allows for feasible PF and OPF solutions.

Step 0: Initialization: Define the 490 power injection scenarios and initial line parameters as discussed above.

Step 1: Solve line upgrade optimization problem: For each power injection scenario, solve the optimization problem (3). This optimization problem attempts to minimize the size of transmission line violations while also limiting generation redispatch away from the assigned power injection schedule. To achieve this, the objective function (3a) is formulated with two terms: (1) a penalty on $\Delta p_{g,s}$, which measures how much generator g is redispatched in scenario s, and (2) a penalty on $\delta_{ij,s}$, which measures the violation of the power flow limit on line ij in scenario s. The factor λ is a trade-off parameter that balances how much we penalize the generation redispatch and line limit violations in the solutions. A smaller value for λ allows for more generation redispatch and leads to fewer line updates. A larger value for λ penalizes generation redispatch more and thus forces more line upgrades. If we do not allow

any redispatch at all, the procedure becomes similar to solving a power flow for each load scenario. Based on testing with several values, we set $\lambda = 0.5$ for our final grid. This seems to give a reasonable trade-off between the number of lines that need upgrades and limiting generation redispatch.

$$\min \quad \lambda \sum_{k \in \mathcal{G}} \Delta P_{g,k}^s + (1 - \lambda) \sum_{(i,j) \in \mathcal{L}} \delta_{ij}^s \tag{3a}$$

$$\text{s.t.} \quad P_{g,k} - \sum_{(i,j) \in \mathcal{L}} \beta_{ij}^k P_{f,ij} = P_{d,k} \quad \forall k \in \mathcal{B} \tag{3b}$$

$$P_{f,ij} = -B_i j(\theta_i - \theta_j) \quad \forall (i,j) \in \mathcal{L} \tag{3c}$$

$$-P_{f,ij}^{max} - \delta_{ij}^s \le P_{f,ij} \le P_{f,ij}^{max} + \delta_{ij}^s \quad \forall (i,j) \in \mathcal{L} \tag{3d}$$

$$P_{g,o}^s - \Delta P_{g,k}^s \le P_{g,k}^s \le P_{g,o}^s + \Delta P_{g,k}^s \quad \forall k \in \mathcal{G} \tag{3e}$$

$$P_{g,k}^{min} \le P_{g,k}^s \le P_{g,k}^{max} \quad \forall k \in \mathcal{G} \tag{3f}$$

$$\delta_{ij}^s \ge 0 \quad \forall (i,j) \in \mathcal{L} \tag{3g}$$

$$\Delta P_{g,k}^s \ge 0 \quad \forall k \in \mathcal{G} \tag{3h}$$

$\mathcal{B}$ and $\mathcal{G}$ are the set of all the buses and generators in the grid respectively. The equality constraints (3b) represent the DC power flow equations, while (3d) represent the relaxed transmission line limits and (3e) represent the generation limits after redispatch. The primary output of this optimization problem is the line limit violations δ_{ij}^s for each line $(i,j) \in \mathcal{L}$ in each scenario $s \in \mathcal{S}$. Note that if the DC power flow solution is feasible for the original power injection scenario s, the optimization problem would set $\delta_{ij}^s = 0$.

Step 2: Identify and upgrade overloaded lines: To identify the overloaded lines, we compute the maximum violation across all scenarios,

$$\delta_{ij} = \max_{s \in \mathcal{S}} \delta_{ij,s} . \tag{4}$$

We then randomly choose a subset of 541 lines to upgrade, corresponding to 5% of the total number of lines in the system[1]. If fewer than 541 lines are overloaded, we upgrade all the lines.

For each line (i, j) that is chosen for an upgrade, we use the following procedure:

(a) Using the table of possible conductor types and MVA ratings for this line, upgrade the type of conductor to the one with the closest higher MVA rating.

(b) If the conductor type has already been upgraded to the highest MVA conductor type and could not be updated using the procedure in (a), increase the number of circuits included in the line by one. Note that the maximum number of allowable circuits per line is 8.

If the line has already been upgraded to have 8 circuits, it is left in its overloaded state until the end of the algorithm. Once the algorithm terminates, we double the ratings of all the lines, which can be understood as changing the conductor type from

---

[1]The reason for upgrading only a subset of lines is that some overload problems may resolve by themselves in the next iteration once the other lines have been upgraded. The random choice of lines to upgrade reflects the fact that the power system has been evolving over a long period of time, and thus is not always built to be optimal for the present day loading conditions.

ACSR to ACSS. By doing this, we (a) remove the remaining overloads and (b) ensure that we account for reactive power flows and losses in the lines, which were so far neglected, when assigning reactive power support.

It should be noted that during this entire process, the number of conductors in each circuit was not changed. This is because, in actual networks, the number of bundled conductors in each circuit is usually limited to two or three, with some exceptions. To get parameters that are as realistic as possible, the number of conductors was thus fixed.

Step 3: Identify and upgrade underutilized lines: A similar set of changes are performed to reduce the ratings of the lines that are under-utilized. Transmission lines with a utilization lower than a given threshold (30% in our case) are classified as underutilized. As done in the case of overloaded lines, a random subset of 541 underutilized lines are chosen. In case the number of underutilized lines is smaller than 541, all such lines are downsized.

For each of the chosen lines (i, j):

(a) Using the table of possible conductor types and MVA ratings, downsize the line by choosing the conductor type that has the closest lower MVA rating

(b) If the conductor type has already been modified to one with the lowest MVA rating, reduce the number circuits in the line by one. Since a line must have at least one circuit, do not decrease the number of circuits in a line once it equal one.

If a line has already been downsized to have just the one circuit of the smallest allowable conductor size and it is still underutilized, do not downsize it further.

Step 4: Check termination criterion: If the number of overloaded lines is below a threshold τ, then the line resizing terminates. Otherwise, we return to solving the optimization problem in Step 1. At the beginning, the threshold τ is set to zero. However, if the algorithm fails to terminate after a certain number of iterations, the threshold is increased every iteration until the algorithm terminates. This is to prevent "cycling" where the algorithm upgrades a set of lines, causing another set of lines to become under-utilized. Correcting these under-utilized lines then causes the first set of lines to become overloaded, thus driving the algorithm into an infinite loop if the threshold τ is not increased.

VI. Assigning Reactive Power Support

The reactive power output of the generators alone is not sufficient to maintain the voltage at each bus within its limits. Thus, to ensure that the grid gives rise to an AC power flow feasible solution where all voltage limits are satisfied, we add reactive power compensation elements to the network using the algorithm described below. Due to the high computational burden associated with solving AC OPF for a network of this size, this algorithm considers only a single power injection scenario corresponding to the maximum load scenario with economic generation dispatch. As mentioned in Section V-B, the thermal limits of the lines are temporarily doubled to ensure that the network is AC OPF feasible and the reactive power flow and network losses are accounted for.
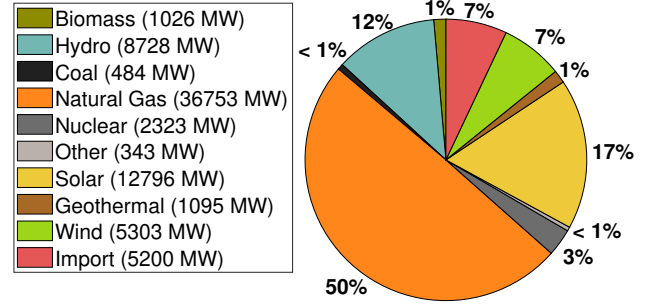

Fig. 6: Percentage of generation from each fuel type.

Step 0: Initialization: Add reactive power compensation in the form of synchronous condensers to all nodes in the network. The initial capacity of the reactive power compensation devices is set to 200 MVAr.

Step 1: Solve AC optimal power flow: For the problem with reactive power compensation installed, we solve a standard AC OPF problem which minimizes generation cost subject to AC PF, generation, transmission and voltage magnitude constraints [24].

Step 2: Remove redundant compensation: Remove reactive power compensation from the 20% of the nodes that currently have reactive power compensation.

Step 3: Check termination criterion: If fewer than 20% of all nodes have reactive power compensation, terminate. Otherwise, go back to Step 1 and resolve the ACOPF.

Step 4: Restoring the line limits: After assigning reactive power support, the thermal limits of lines with a utilization of less than 50% are brought back to their original values before they were temporarily doubled. This results in around 2.54% of lines with limits that are doubled, corresponding to an upgrade from an ACSR conductor to an ACSS conductor, as described in Section V-A3.

VII. Grid Metrics and Evaluation

A brief overview of the network structure is given in Table I. The created synthetic network has 8848 buses, out of which 1,483 have reactive power support in the form of synchronous condensers. There are 10,140 transmission lines and 663 transformers in the grid. The system has 2461 load buses with a peak load of 44,001.4 MW and 2,229 generators with a total capacity of 74,052 MW. Generation capacities for different fuel type is shown in Fig. 6.

An important metric to evaluate synthetic networks is the node degree distribution, which captures the frequency of the node degree (number of lines connected to the substation) of each substation. Fig. 7 shows the node degree distribution for our synthetic grid. As expected, the node degree distribution of our network follows a downward trend and agrees closely with the trends followed by real networks [19].

Characteristics of the network branches, shown in Table II, follow trends similar to those of real grids presented in [19].
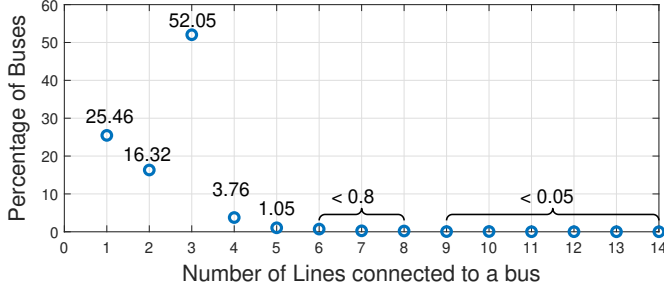
Fig. 7: Node degree distribution for the network.

TABLE I: Network data

Metric	Number
Buses	8848
Transformers	663
Transmission lines	10140
Generators	2229
Buses with reactive power compensation	1483
Load Buses	2461

Any statistical comparison of the synthetic network and real grids will have shortcomings due to the fact that there are only three samples of real world networks in the United States.

We evaluated the operation of the CATS grid across a year of hourly renewable generation and load scenarios. The renewable generation data comes from 2019 aggregate production data from CAISO [13]. As described in Section IV-2, for each hour, the capacity of each renewable generator is scaled according to its rating and the aggregate renewable production. For the load scenarios, we use a full year of the previously referenced hourly load data (see Section II). For each hour, we solve a DC and an AC OPF. Out of the 8,760 hourly time steps, each scenario is feasible and is solved to a locally optimal solution.

We then evaluate aspects of the power flow solutions, including feasibility, line loading, generation dispatch, and curtailment. Transmission line congestion is shown in Table III. Out of 10,140 transmission lines in the network, there are on average 11.4 lines that are operating at their maximum capacity in the DC OPF solutions, while 2.25 on average are operating at their capacity on the AC-OPF solutions. The highest congestion level in the DC OPF solutions occurs on Aug 14 at 7pm and Sep 4 at 7pm, when 26 lines are operating at their capacity. In the AC Scenario, a maximum of 30 lines are simultaneously at their peak capacity and this occurs Aug 14 at 7pm. Overall, there is a small level of congestion in this network under most scenarios, but the congestion is limited and the grid generally operates with economic efficiency.

TABLE II: Characteristics of the network branches

Voltage level (kV)	GVA-miles	Total length (miles)	Percent of lines
66	957.11	12696.34	0.6485
115	1755.71	8076.1	0.2170
230	5741.30	8256.5	0.1225
500	10022.20	4637.2	0.0120

TABLE III: Congested transmission lines

PowerFlow	Mean	Median	Maximum	Minimum
DC-OPF	11.4	12	26	0
AC-OPF	8.62	9	30	0

TABLE IV: Hourly cost [millions]

PowerFlow	Mean	Median	Maximum	Minimum
DC-OPF	$30.07	$22.11	$93.22	$0.45
AC-OPF	$12.73	$6.85	$48.79	$0.00

The cost of operation in shown in Table IV. The operating cost is on average $30 million USD per hour of operation, but the maximum operating cost approaches $100 million and the minimum approaches $1 million. The AC OPF solution typically costs 20% more than the DC OPF solution.

The hourly generation is shown in Table V. The DC OPF problem does not compute losses, and so the generation is equal to demand. The hourly generation ranges from 15,352 MWs to 39,950 MWs, with a average of 22,714 MWs of generation. The AC solution requries approximately 5% more generation to account for losses in the network.

Finally, in Table VI we show the curtailment of wind and solar in the grid. The median curtailment for all hours of the year is 0 MW, and on average only 26.5 MW of solar or wind is curtailed in the DC solutions, or 8.5 MW in the AC solutions. Only 241 out of 8760 hours of the year have curtailment over 100 MW in the DC solutions, and 102 hours have curtailment over 100 MW in the AC solutions. It is important to note that this is not reflective of much higher curtailment levels in the actual CAISO system. We suspect that the low curtailment is due in part to the fact that we scale the wind and solar capacity by the dispatched power of the CAISO market, not the total available wind or solar at a given hour. An area of future work is to improve the accuracy of the renewable generation profiles.

Generation profiles of the AC solutions are shown in Fig. 9 for a winter profile (January 7 through 13) (Fig. 9a) and a summer profile (August 1 through 7) (Fig. 9b). The generation output in the summer is much higher and has a larger ramp rate, as is expected for California. This is especially pronounced for natural gas (orange) after solar output (yellow) drops after sunset. We also note that the daily production period for solar energy is shorter in the winter than the summer, since there are fewer sunlight hours in the winter.

VIII. USING THIS SYNTHETIC GRID

The CATS grid model can be used as a test case for power systems research. It is particularly valuable for use in geo-referenced applications, such as those relating to weather, climate change, topography, political boundaries, socio-economic considerations, and more. The CATS is available in a GitHub repository [15]. The MATPOWER (.m) file is a standard format for storing power system component information and is commonly used as an input for a number of power system analysis tools, such as PowerModels.jl. The GIS data are in the GEOJSON format and can be opened in GIS mapping software, like ArcGIS Pro. Additional data files are available in an external folder that is referenced from the repository.
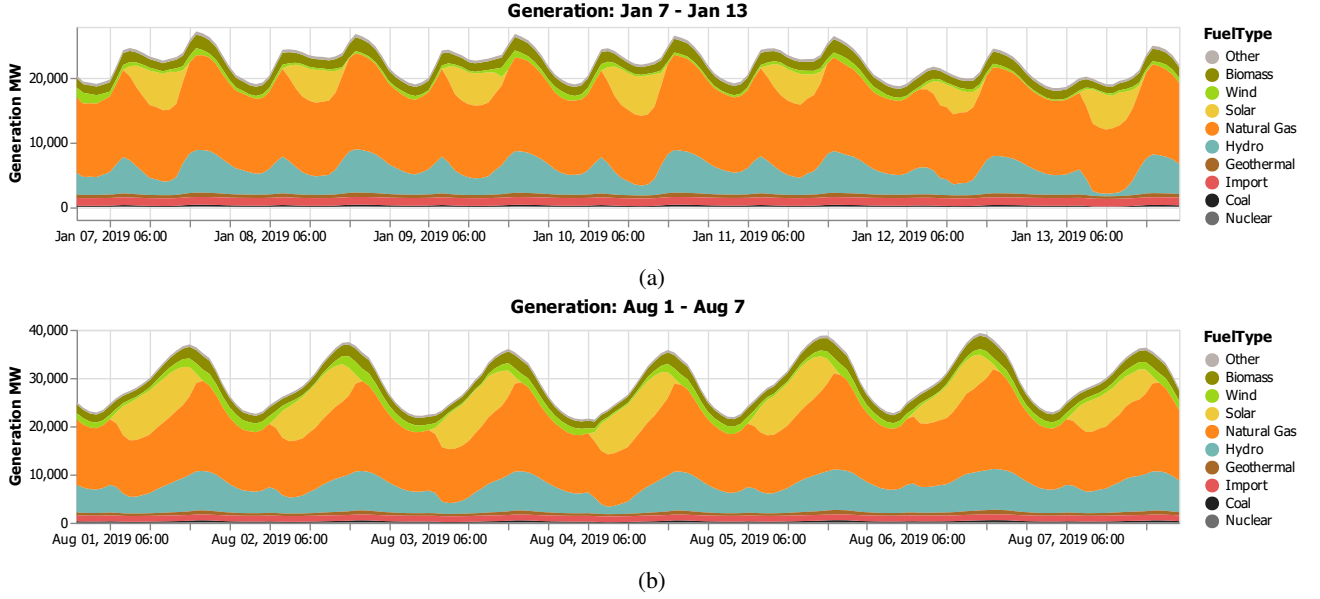
Generation: Jan 7 - Jan 13

(a)

Generation: Aug 1 - Aug 7

(b)

Fig. 8: Generation dispatch from the AC solutions for a) Jan 7 through Jan 13. and b) Aug 1 through Aug. 7.

Generation: Jan 7 - Jan 13

(a)

Generation: Aug 1 - Aug 7

(b)
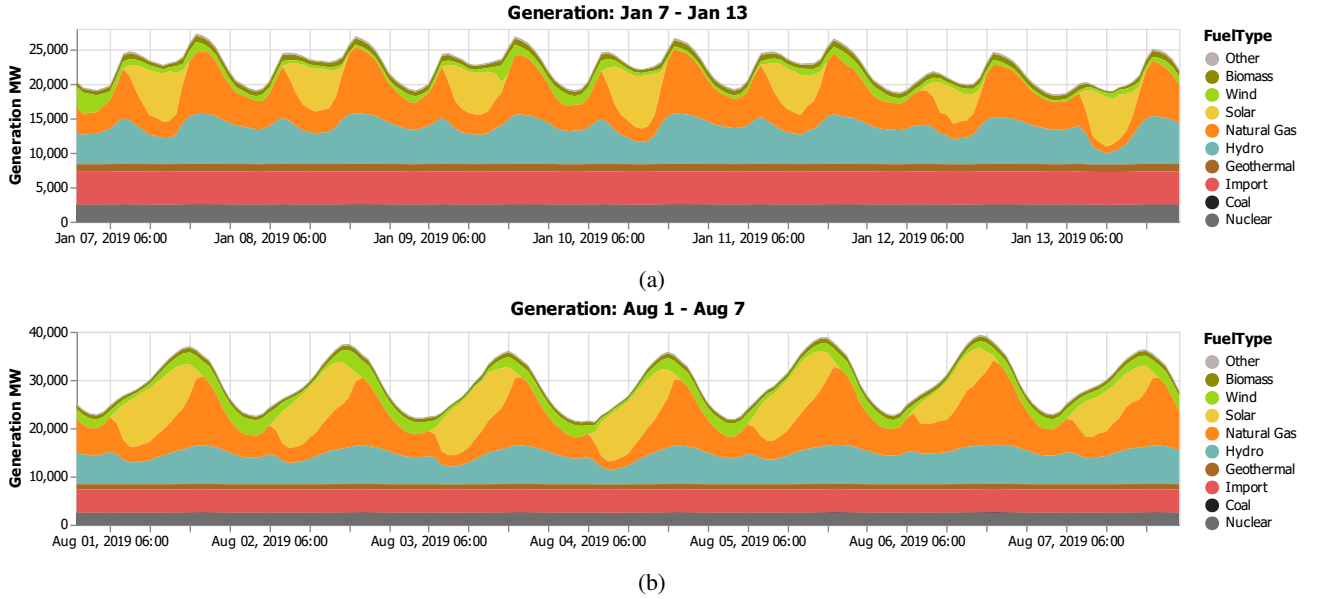
Fig. 9: Old figures for comparison

TABLE V: Hourly generation [MW]

PowerFlow	Mean	Median	Maximum	Minimum
DC-OPF	22714	21942	39950	15352
AC-OPF	23928	23137	42060	15998

TABLE VI: Total curtailment [MW]

PowerFlow	Mean	Median	Maximum	Minimum
DC-OPF	27.2	0.0	2797	0.0
AC-OPF	1385	1220	5930	0.0

IX. CONCLUSION

The main contribution of this work is the California Test System (CATS), a synthetic grid that represents the real locations of California's electric infrastructure with invented inter-connections and parameters. To the authors' best knowledge, this is the first publicly available test system to accurately reflect the geography of California's transmission system.

It is important to note that the final network is merely an approximation of California's transmission system. While the locations and paths of the components were not significantly modified, the connections and parameters are invented. This is important for maintaining security, but it also means that any results produced using this grid do not necessarily reflect the results that would be produced from California's actual grid.

Another limitation is the connectivity of the final system. In the current version of this test case, not all of the buses are included in the largest connected network. However, we believe that this is reasonable because it is consistent with

other publicly available synthetic grids, such as the Western Electricity Coordinating Council grid model.

The authors have identified several areas for future development. In particular, we can improve the realism of the renewable generation time-series data by incorporating weather data with geographic variability.

REFERENCES

[1] "Critical Energy/Electric Infrastructure Information (CEII)," Federal Energy Regulatory Commission, 2021. [Online]. Available: www.ferc.gov/ceii

[2] "IEEE test cases," IEEE Power and Energy Society. [Online]. Available: site.ieee.org/pes-tccwg/links-to-test-cases

[3] "Texas A&M University Electric Grid Datasets." [Online]. Available: electricgrids.engr.tamu.edu

[4] "University of Wisconsin-Madison Synthetic Electric Network Models." [Online]. Available: github.com/WISPO-POP/SyntheticElectricNetworkModels

[5] C. Barrows, A. Bloom, A. Ehlen, J. Ikäheimo, J. Jorgenson, D. Krishnamurthy, J. Lau, B. McBennett, M. O'Connell, and E. Preston, "The IEEE Reliability Test System: A proposed 2019 update," *IEEE Trans. on Power Systems*, vol. 35, no. 1, pp. 119–127, 2019.

[6] S. Babaeinejadsarookolaee, A. Birchfield, R. D. Christie, C. Coffrin, C. DeMarco, R. Diao, M. Ferris, S. Fliscounakis, S. Greene, R. Huang, C. Josz, R. Korab, B. Lesieutre, J. Maeght, D. K. Molzahn, T. J. Overbye, P. Panciatici, B. Park, J. Snodgrass, and R. Zimmerman, "The power grid library for benchmarking AC optimal power flow algorithms," *arXiv preprint arXiv:1908.02788*, 2019.

[7] 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. on Power Systems*, vol. 32, no. 4, pp. 3258–3265, 2017. [Online]. Available: https://dx.doi.org/10.1109/TPWRS.2016.2616385

[8] A. Birchfield, E. Schweitzer, M. Athari, T. Xu, T. 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. [Online]. Available: www.mdpi.com/1996-1073/10/8/1233/htm

[9] S. Babaeinejadsarookolaee, J. Snodgrass, S. Acharya, S. Greene, B. Lesieutre, and C. L. DeMarco, "Comparison of real and synthetic network models of the western United States with respect to new realism measures," in *2021 IEEE Power and Energy Conference at Illinois (PECI)*. IEEE, 2021, Conference Proceedings, pp. 1–8.

[10] N. Rhodes, L. Ntaimo, and L. A. Roald, "Balancing wildfire risk and power outages through optimized power shut-offs," *IEEE Trans. on Power Systems*, vol. 36, pp. 3118–3128, 2021.

[11] "California electric transmission lines," California Energy Commission, September 2020. [Online]. Available: gis.data.ca.gov/datasets/CAEnergy::california-electric-transmission-lines/about

[12] "Form EIA-860 detailed data with previous form data (EIA-860A/860B)," Energy Information Agency, 2019. [Online]. Available: www.eia.gov/electricity/data/eia860

[13] "Production and curtailment data," California Independent Systems Operator (CAISO), 2019. [Online]. Available: www.caiso.com/informed/Pages/ManagingOversupply.aspx

[14] "Form 1 - Electric Utility Annual Report," Federal Energy Regulatory Commission, Mar 2022. [Online]. Available: www.ferc.gov/general-information-0/electric-industry-forms/form-1-electric-utility-annual-report

[15] "CATS-CaliforniaTestSystem," Oct. 2022. [Online]. Available: github.com/WISPO-POP/CATS-CaliforniaTestSystem

[16] S. Babaeinejadsarookolaee, Tech. Rep., 2018. [Online]. Available: https://tinyurl.com/48b68t2s

[17] H. Zaininger, A. Wood, H. Clark, T. Laskowski, and J. Burns, "Synthetic electric utility systems for evaluating advanced technologies. final report," Power Technologies, Inc., Tech. Rep., 1977.

[18] L. Wang, J. McCalley, C. L. DeMarco, F. Akhavizadegan, A. Venkatraman, and J. Snodgrass, "Development of expansion planning methods and tools for handling uncertainty," Power Systems Engineering Research Center, Tech. Rep., May 2020.

[19] J. M. Snodgrass, "Tractable algorithms for constructing electric power network models," Ph.D. dissertation, The University of Wisconsin-Madison, 2021.

[20] "Form no. 715 - Annual Transmission Planning and Evaluation Report," Federal Energy Regulatory Commission. [Online]. Available: www.ferc.gov/industries-data/electric/electric-industry-forms/form-no-715-annual-transmission-planning-and-evaluation-report

[21] Southwire Co., "ACSR aluminum conductor. steel reinforced. bare." Tech. Rep., 2012. [Online]. Available: https://tinyurl.com/54yuhkb6

[22] J. J. LaForest, *Transmission line reference book, 345 kV and above*, second edition. ed. Palo Alto, Calif.: General Electric Company. Large Transformer Division and Energy Systems and Technology Divisions Electric Power Research Institute., 1982.

[23] Southwire Co., "ACSS aluminum conductor, steel supported. bare." Tech. Rep., 2017. [Online]. Available: https://tinyurl.com/yc5e4wcd

[24] C. Coffrin, R. Bent, K. Sundar, Y. Ng, and M. Lubin, "PowerModels.jl: An open-source framework for exploring power flow formulations," in *2018 Power Systems Computation Conference (PSCC)*. IEEE, 2018, Conference Proceedings, pp. 1–8.