Open Power System Datasets and Open Simulation Engines: A Survey Towards Machine Learning Applications

Ignacio Aravena, *Member, IEEE,* Chih-Che Sun, *Senior Member, IEEE,* Ranyu Shi, *Student Member, IEEE,* Subir Majumder, *Member, IEEE,* Weihang Yan, *Member, IEEE,* Jhi-Young Joo, *Member, IEEE,* Le Xie, *Fellow, IEEE,* Jiyu Wang, *Member, IEEE*

***Abstract*—A major factor behind the success of machine learning (ML) models is the availability and accessibility of large, well-organized datasets for training and benchmarking. In regards to the availability of the power grid datasets, there are two major challenges: (i) real-world data is often restricted by regulatory constraints, privacy reasons, or national security concerns, making it difficult to obtain, and (ii) synthetic datasets, which are created to address these limitations, are often generated using specialized power grid simulation tools that are not easily accessible to the broader ML community. Therefore, establishing a comprehensive and standardized database that serves as a central repository for publicly available power system datasets is essential to facilitate the development of ML tools in the power engineering field. This paper addresses these challenges by reviewing the current landscape of publicly accessible datasets, including open-source power grid network data, detailed machine models, consumer demand profiles, renewable generation data, and inverter models. We also examine open-source power system simulators, which are crucial for generating high-quality, high-fidelity power grid datasets. This survey aims to provide a foundation for overcoming data scarcity and to guide the creation of a unified database that can support the development of ML-based tools for power grid engineering.**

***Index Terms*—IEEE, IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

Machine learning (ML) techniques are transforming operations across industries, particularly where the abundance of real data has enabled large offline training and validation. In contrast, the electricity sector faces significant challenges due to limited access to real system data, which is often kept confidential for national security and competitive market access reasons. Despite past and ongoing efforts to create high-quality synthetic datasets (*e.g.*, projects within ARPA-E's Grid Data Program [1] and recent datasets derived from them [2]), publicly available data remains scarce and typically tailored for specific applications (e.g., planning, operations, or transient simulations) because they are generated with a certain application in mind. This leads to four primary issues with existing open datasets:

1) they are *incomplete*, only include technical details needed for certain types of studies (e.g., power flow analysis), lacking comprehensive information;
2) they are *modified* over time by the original authors or other researchers, leading to divergent and sometimes incompatible versions, creating inconsistent branches of the same dataset;
3) they are often *biased*, as these datasets are constructed to demonstrate proof-of-concept for specific analytical techniques and are not fully representative of real systems (e.g., transmission MVA ratings, modified IEEE test cases); and
4) they are released in *specialized formats* for specific applications, making them incompatible with other applications and not maintained over time.

Each of the listed issues leads to detrimental consequences. Incomplete datasets prevent researchers from conducting holistic studies (*e.g.*, verifying dynamic stability for expansion planning decisions) hindering progress in developing methodologies that reflect the full complexity of industrial workflows, which routinely utilize real and complete system datasets. Incomplete datasets also lead to the second issue, divergent versions of the same dataset, which can lead to inaccuracies when performing comparisons or validations across studies, as cumulative changes can produce vastly different results even for identical use cases. For example, a power flow solved with a dataset with original impedances may yield very different results compared to solving with adjusted impedances. The third issue, system parameter biases, can lead to incorrect assessments of novel methods. For example, symmetry in generator parameters can be of little importance for a reliability study, but for unit commitment it can lead to detrimental performance, even when perfect generator symmetry is never present in the real world. This type of issue can mislead researchers to focus and design solutions for challenges not present in real systems. The fourth data format issue is less pervasive because of the development of standardized data representation for power systems, the Common Information

I. Aravena, R. Sun, and J.-Y. Joo are with the Computational Engineering Division of Lawrence Livermore National Laboratory, Livermore, California, USA. Email: {aravenasolis1,joo3,sun31}@llnl.gov

L. Xie is with the Harvard John A. Paulson School of Engineering and Applied Sciences, Boston, Massachusetts. Email: xie@seas.harvard.edu

R. Shi, and S. Majumder are with the Department of Electrical and Computing Engineering of Texas A&M University College of Engineering, Texas, USA. Email: {le.xie,siennashi,subir.majuder}@tamu.edu.

W. Yan, and J. Wang are with National Renewable Energy Laboratory, Colorado, USA. Email: {le.xie,siennashi,subir.majumder}@tamu.edu.

Manuscript received July xx, 2023; revised YYY zz, 2023. Document released as LLNL-JRNL-868871

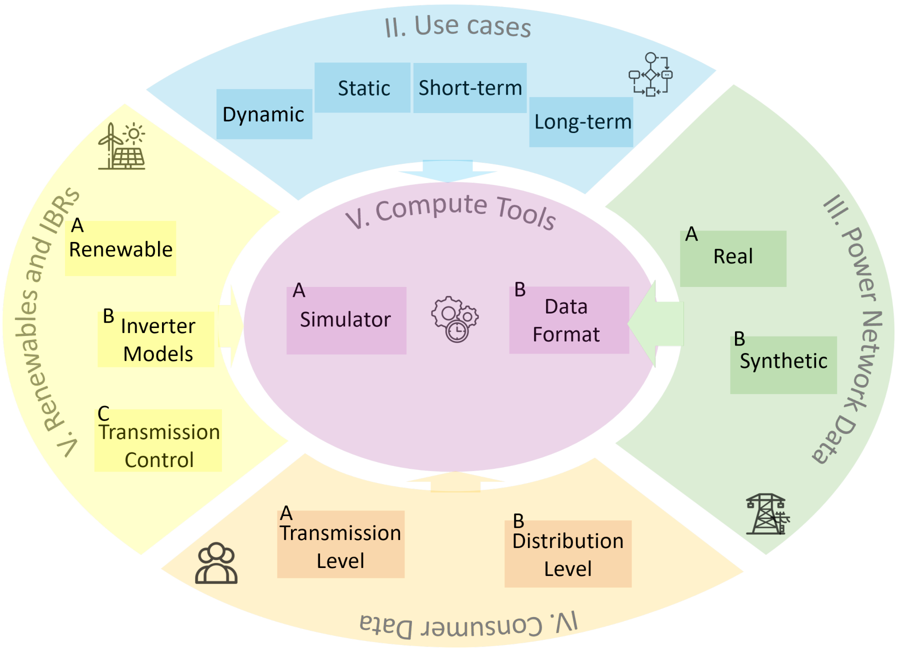

Fig. 1. Overview of key publicly available datasets and tools in power systems research. The figure serves as a guide for navigating through different sections: Section II: Use Cases, Section III: Power Network Data, Section IV: Consumer Data, and Section V: Renewables and Inverter-Based Resources (IBRs), and Section VI: Compute Tools (simulators and data formats).

Model (CIM); they pose significant entry barriers for new researchers, especially those outside the power systems field.

Additionally, while power systems datasets are scarce, input/output datasets representing the physical phenomenon for ML models to learn (e.g., *input*: perturbation, *output*: dynamic response) are almost nonexistent or do not contain enough information around the frontier of interest for particular cases study (e.g., balanced stable and unstable dynamic responses [3]), making the offline training of certain ML models (e.g., classifiers) very challenging. This also makes simulators indispensable as a part of the high-quality data generation process for ML-based tool development in power engineering. This leads to a further challenge even when a traditional power system dataset is made accessible to ML researchers, both open and commercial analytical tools for power systems appear very cryptic to non-expert users; it is nontrivial to discern which are the most important use cases or models that could be aided by certain ML technology from all other possible simulations, and the process of setting up and running simulations (i.e., finding initial conditions, simulation horizon, and other simulation settings) is also a significant challenge.

This survey paper aims to lower these entry barriers to power systems datasets by summarizing the currently available public datasets, models, and simulation tools in power systems domains that are necessary to generate a database of synthetic data to address the accessibility challenges. Here our objectives are threefold: 1) to *assist* researchers in both power systems engineering and ML domains in identifying suitable datasets, models, and simulation engines for various purposes, 2) to *organize* and present these resources in a manner that is accessible to users outside the power systems field, making it easier to understand their functionalities and primary use cases, and 3) to *promote* data science applications to these datasets and models by fostering interdisciplinary collaboration between the power systems and machine learning communities. In this regard, Fig. 1 provides a structured roadmap for researchers, illustrating how to navigate through different sections to locate specific aspects of power grid data and tools relevant to their research focus.

Here, Section II presents concise descriptions of the key use cases for power system simulations, highlighting that power grid research involves multiple timelines. Section III catalogs open datasets for bulk power network infrastructure—including transmission lines, transformers, generators, switches, and their interconnections—covering both real-world and synthetic systems. Section IV lists existing data and models for consumers, with a particular emphasis on flexible consumers, who are playing an increasingly significant role in modern power grids. Section V summarizes models for renewable and inverter-based resources, whose fast integration has significantly impacted power grid operations in recent decades. Section VI lists existing open-source simulators with their use cases, as well as publicly available, common formats for power system data. The lists in Sections III through VI are compiled to the best of the authors' knowledge, but there may be unintended omissions. Finally, Section VII offers recommendations for enhancing power grid data quality and accessibility.

II. Power System Simulation Use Cases

There exist multiple types of simulations involving bulk power systems. These range from detailed simulation of individual components (e.g., wave propagation on transmission lines, switching models for inverters, Park equations for electrical drives); passing by simulations of subsets of the grid (e.g., interactions between reactive controllers); whole interconnected systems simulations (e.g., power flow and contingency analysis); up to interdependent systems simulations (e.g., hydro-thermal coordination). In this work we focus in the third class, simulations of interconnected power grids, within which we distinguish 4 important groups of use cases:

1) **Dynamic simulation** represents how the system state evolves in the instants following an abrupt change or perturbation. Typical use cases include short circuit analysis, switching (topology changes) and line tripping simulations, forced outage simulations, and stability analysis.
2) **Steady-state single-snapshot simulation** corresponds to calculations for all the quantities in the system for a given snapshot with certain boundary or initial conditions. Among use cases here we find power flow calculations, (steady-state) contingency analysis, and optimal power flow calculations.
3) **Short-term operation simulation** involves simulating the mechanisms used to schedule the operation of the power grid within a few hours to a few days. Use cases include day-ahead, intraday, and real-time market clearing; unit commitment and economic dispatch; emergency preparation and response; and power system blackstart and restoration.
4) **Long-term operation simulation** involves simulating the power grid for anticipated future conditions of its long-term operation and needs. Use cases within this group include maintenance scheduling, reliability studies, and generation and transmission expansion planning.

Executing each of the use cases above requires a different set of technical or economical parameters for power grid components. The total number of necessary parameters to execute all possible use cases listed above for a single component can easily exceed 30 (e.g., complete transformer model with sequence data), whereas for a single use case even 2 parameter might be enough (e.g., transformer nominal capacity and series reactance for unit commitment). This dramatic difference explains why most synthetic datasets for power grids contain information only for a few use cases, typically within the same group, as the effort required to generate (sensible/validated) complete sets of parameters is a demanding tasks on its own [1].

III. Open Power Network Datasets

A. Open real network datasets

The availability of open-sourced power system model datasets has significantly advanced the field of power systems research and analysis. These datasets provide researchers and practitioners with access to comprehensive, real-world data that can be used to model, simulate, and optimize power system operations. By leveraging open data, stakeholders can collaborate more effectively, drive innovation, and develop solutions to complex challenges facing modern power grids. Four well-known open datasets from the real-world are introduced in this section.

1) **Chilean power grid data** [4]: This open dataset offers extensive and granular details on all components of the Chilean electrical grid, providing a robust foundation for comprehensive simulations and analysis. The dataset features dynamic and sequence data, which are crucial for understanding the temporal behaviors and event sequences within the power network. Dynamic data encompasses time-series information on power flows, voltages, and frequencies, enabling the study of system stability, transient events, and load variations. Sequence data provides insights into the order and timing of specific events, such as faults, switching operations, and protective relay actions, which are essential for fault analysis and the development of preventive measures. The data is only accessible in Spanish.
2) **Colombian power grid data** [5]: The dataset primarily focuses on generator characteristics. This information is essential for analyzing the country's generation capacity and planning grid expansion. However, the dataset's limitation in not providing detailed data on other grid components, such as transmission lines and substations, may restrict its applicability for comprehensive power system studies. Nevertheless, it remains a valuable resource for understanding the generation infrastructure in Colombia. The data is only accessible in Spanish.
3) **Open energy modeling (openmod)** [6]: This is a community-driven effort involving energy system modelers from universities and research institutions. This initiative advocates for the adoption of open-source software and open data in energy system modeling, with a focus on research and policy guidance. Openmod compiles both real and synthetic datasets, providing documentation on various open-source energy models and exploring practical and theoretical challenges related to their creation and use. Transmission network models from multiple regions around the world are available in the database, including primarily Europe, Australia, United States, and other regions. Additionally, a few non-region-specific power system models are also presented in the openmod database.
4) **Open Power System Data** [7], [8]: The development of this data platform began in 2014 with the aim of addressing the inefficiencies and challenges faced by researchers in collecting and processing the vast amounts of data required for power system modeling. The platform is designed to provide European power system data in five packages: (1) conventional power plants, (2) national generation capacity, (3) renewable power plants, (4) time series, and (5) weather data.

B. Open synthetic network datasets

Synthetic power system models offer a critical advantage in power system research by providing detailed, scalable, and customizable network models that can be freely accessed and utilized. Unlike real-world data, which may be restricted due to privacy or proprietary concerns, synthetic datasets can be designed to represent a wide range of scenarios and conditions. This flexibility allows researchers to explore and evaluate different network configurations, operational strategies, and system behaviors, facilitating the development of more robust and versatile power system models and solutions. In this section, several open synthetic datasets are reviewed.

1) **NREL's reduced Western electricity coordinating council (WECC) system** [9], [10]: The transition to inverter-based resources (IBR) and the retirement of synchronous generators present new challenges for bulk power system planning and operation. To address this, a dynamic model of a reduced WECC system has been developed by NREL. This simplified 240-bus WECC model integrates with existing scheduling models to support both scheduling and dynamic simulations. Reflecting the 2018 generation mix of the Western Interconnection, the model is validated with field frequency data from FNET/GridEye [11] and accurately represents key inter-area oscillations in WECC .
2) **Texas A&M electric grid test cases** [12], [13]: This power system model repository offers synthetic power grid models designed to address the challenges of modern power grids, supporting both research and education. The repository includes test cases of varying sizes, ranging from 9 to 82,000 buses, with each test case available in multiple formats, including PowerWorld/PowerWorld DS, Matpower, PSS/E, and PSLF.
3) **PNNL's DR POWER platform** [14]: This repository is supported by the ARPA-E GRID DATA Program through the DR POWER project. Its primary objective is to establish, maintain, and develop open-access power grid models and scenarios, commonly referred to as datasets. The repository offers a variety of models for power system simulations, encompassing weather, transmission and distribution (T&D) systems, economics and markets, as well as generation.
4) **Power Grid Lib (PGLib)** [15]: This benchmark library is overseen by the IEEE Power and Energy Society (PES) Task Force on Benchmarks for Validation of Emerging Power System Algorithms. Two repositories are introduced as follows:
   pglib-OPF is specifically developed to assess a well-established version of the AC optimal power flow problem. The accompanying introductory video and comprehensive report outline the library's motivations and objectives. These cases are particularly crafted to benchmark algorithms that address the non-convex nonlinear program.
   pglib-UC is intended to assess a widely recognized version of the unit commitment problem. These cases are specifically created to benchmark algorithms that solve the mixed-integer linear program.
5) **IEEE Test Case Coordination Working Group tests** [16]: This work highlights a variety of power system modeling test cases designed for research and education in areas such as voltage stability, cascading failures, and renewable energy integration. These test cases include widely-used IEEE models, such as distribution feeders and voltage stability test systems, along with newer contributions from the ARPA-E GRID DATA program and other initiatives. Additional resources like MATPOWER and WECC test cases further enhance the study of power flow and market dynamics. Collectively, these tools offer comprehensive support for advancing power system modeling and analysis.
6) **Open Infrastructure Map, Open Grid Map** [17]: This project provides a view of the world's infrastructure based on data from OpenStreetMap. The data includes detailed location of power lines, substations, and power generation plants, as well as basic technical information for the above (e.g., capacity of generators), but does not contain electrical parameters of the transmission system that would permit simulation. Nevertheless, the topological information has been used to produce synthetic electrical parameters based on typical values [18], which resulted in a detailed synthetic dataset for Germany. Future development of these map-based generation methodologies might lead to synthetic power grid data with world-wide coverage.

While synthetic power system models offer valuable flexibility and accessibility, they also have limitations compared to real datasets. One of the key challenges is the inherent lack of real-world complexity and variability. Synthetic datasets are typically designed to represent a wide range of homogeneous scenarios, but they may not capture the nuanced behaviors, unexpected anomalies, or specific operational challenges found in actual power systems. Additionally, because synthetic data is generated based on assumptions and models, it may not fully reflect the impact of neglected external factors such as weather conditions, equipment aging, or human interventions. These limitations highlight the importance of using real datasets to validate synthetic models and motivate the need for continued research into improving the fidelity of synthetic power system models.

IV. Open Flexible Consumer Datasets

Technological advancements and the growing integration of renewable energy sources and storage systems are enabling consumers to play a more active role in managing their energy consumption. According to the U.S. Federal Energy Regulatory Commission's 2020 Staff Report, the deployment of Advanced Metering Infrastructure (AMI) in the residential sector expanded from 6.7 million meters in 2007 to 88 million in 2018, significantly enhancing the potential for demand flexibility [19]. However, the extent of this flexibility is influenced by the specific demand response mechanisms available and the preferences of different customer segments. From 2016 to 2021, the U.S. Energy Information Administration reported that true peak demand savings grew from 34% to 50% for residential customers, from 32% to 39% for commercial customers, and from 32% to 44% for industrial customers, reflecting variations across sectors [20]. Figure 2 illustrates the total actual peak demand savings by state for 2021.

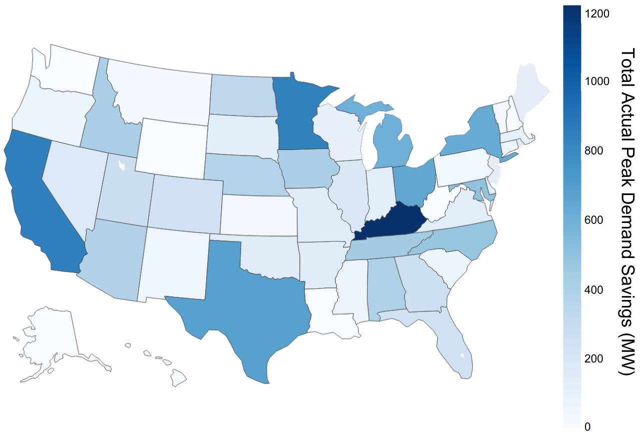

Fig. 2. 2021 Total Actual Peak Demand Savings in Each State

In this section, we will review publicly available datasets that document consumer contributions to system flexibility, and in this regard, we categorize consumers into two groups based on their operational characteristics: (A) transmission-level consumers—such as data centers, which can draw upwards of 900 MW—and (B) aggregated distribution-level consumers.

A. *Transmission-level consumers*

At the wholesale level, demand response resource participation increased from 30,587.7 MW in 2020 to 32,421.0 MW in 2021 across all seven ISOs/RTOs in the continental United States [19]. Recent initiatives by grid operators to engage large-scale consumers underscore the effectiveness of demand flexibility strategies. For instance, the *Flex Alert* program, implemented by the California Independent System Operator (CAISO) during the September 2022 heatwave, successfully reduced demand by 2 GW [21]. However, non-emergency efforts to leverage consumer flexibility have yielded mixed results, influenced by various factors. In the following we provide some of the datasets for large customers offering flexibility.

1) **ERCOT ancillary service data** [22]: The Electric Reliability Council of Texas (ERCOT) currently deploys five types of ancillary services: Regulation services, responsive reserve, non-spinning responsive reserve, and ERCOT contingency reserve service, as a part of their total demand response program. The data for ERCOT ancillary services can be requested through the ERCOT public portal through their API.
2) **Demand response across the continental US for 2006** [23]: The dataset is designed to evaluate hourly demand response availability across the continental U.S., segmented by balancing authority, end-use category, and grid function. It provides a detailed overview of demand response potential by examining 14 end-use categories, covering residential, commercial, industrial, and municipal sectors. Furthermore, it considers five essential bulk power system services: regulation reserve, ramping reserve (flexibility), contingency reserve, energy provision, and capacity. Demand response availability is estimated by matching the specific physical requirements of these grid services with the corresponding load profiles for each end-use category.

B. *Aggregate distribution-level consumers*

According to the U.S. Federal Energy Management Program [24], the scope of an aggregated demand response program often depends on the type of dynamic rates employed, ranging from time-varying pricing to real-time pricing. All of these programs are categorized as “non-dispatchable,” meaning that customers respond voluntarily to price signals rather than being directly controlled. Therefore, the demand profiles could provide valuable insights into consumer behavior and are typically classified as sensitive information. Therefore, a consistent and comprehensive dataset is lacking. While federal agencies such as FERC and NERC collect demand response data through initiatives like advanced metering surveys and the Demand Response Availability Data System (DADS), these datasets are not publicly accessible and typically exclude smaller resources. Currently, the only reliable open-source datasets available are from the UK Power Networks-led Low Carbon London project [25] and residential data provided by Pecan Street [26].

1) **CityLearn** [27], [28]: CityLearn is an open-source platform for modeling power system loads in urban environments, focusing on building energy coordination and demand response. It provides a simulation environment to test and benchmark reinforcement learning (RL) algorithms for optimizing energy use across city districts through coordinated control of storage and heating systems. This flexibility enables standardized comparisons of RL agents across different scenarios and strategies.
2) **Pecan Street data** [26]: provides access to Dataport, one of the largest publicly available repositories of residential energy usage data globally. This data portal includes high-resolution data on electricity consumption, solar generation, electric vehicle charging patterns, and water usage at the household level. By offering detailed insights into energy use and behavior, this expansive resource supports a wide range of research areas, including demand response, smart grid analytics, and energy efficiency.
3) **Low Carbon London** [25]: provides an extensive dataset from energy consumption readings of 5,567 London households collected between November 2011 and February 2014. This dataset, comprising approximately 167 million half-hourly readings, offers detailed information such as energy consumption in kWh, unique household identifiers, and timestamps. The data encompasses two groups: approximately 1,100 households exposed to Dynamic Time of Use (dToU) pricing throughout 2013, receiving varying price signals (High, Low, or Normal) to manage energy use and testing responses to high renewable generation and grid stress.

The remaining 4,500 households were on a flat-rate tariff. This comprehensive dataset supports advancements in understanding demand response and grid management strategies.

V. Open Renewable and Inverter-based Resource Datasets

As renewable energy integration is growing rapidly, it is becoming significantly critical to recognize and integrate the characteristics of the renewable energy resources into grid planning, operation, and system health monitoring. There are two main features of the renewable energy resources which are important to power system simulations:

1) **Generation intermittency**. Because the generation of the renewable energy resources such as solar and wind are largely impacted by weather, their generation are fluctuating along with the weather conditions.
2) **Inverter-based resources**. Unlike the traditional power plants with synchronous generators, renewable energy resources are inverter-based resources that do not synchronize with the system frequency by physical coupling automatically. This can impact the power system which are originally designed with power plants with synchronous machines.

It is crucial to effectively integrate the above two main features into power system planning, operation, and health monitoring. In this section, we'll review the open datasets which are available to be leveraged by the stakeholders to analyze and integrate the two main features.

A. Open primary renewable energy resource databases

1) **National Solar Radiation Database (NSRDB)** [29]: NSRDB is a serially complete collection of hourly and half-hourly values of the three most common measurements of solar radiation—global horizontal, direct normal, and diffuse horizontal irradiance—and meteorological data, also includes wind direction, wind speed data. Users can input the desired address, or directly select a point or a zone on the map to find the nearest available dataset.
2) **NASA The POWER Project** [30]: NASA prediction of worldwide energy resources provide solar and meteorological data sets from NASA research for support of renewable energy, building energy efficiency and agricultural needs. This dataset includes hourly solar irradiance data and wind speed data. Users can input geographic information, or select a point on the map to find the nearest available dataset.
3) **Total Weather Data** [31]: Total Weather Data includes hourly temperature, humidity, snow, precipitation, wind, cloud cover, and solar radiation data. Users can input a desired location to find the nearest available dataset.

B. Open renewable generator and inverter models

To better understand the dynamic behavior of power electronics-based converter, electromagnetic transient (EMT) simulation is necessary to under the stability properties and generation characteristics of inverter-based renewable energy resources. Hence, the majority of the available models conducting EMT simulations are based on graphical and block diagram-based software, such as Matlab/Simulink, PSCAD, Digsilent, etc. where user-defined controls can be implemented.

A number of implementations of IBRs, including Type III/IV wind turbine generators, 2/3 level PV inverters, are available as demo models in these software environments. These models are well validated and can be selected to simulate as average models or switching models, depending on the requirements of simulation studies. A list of existing average models for Type III wind energy simulations is provided in Table I.

C. Open models for HVDC lines, grid-scale batteries, and other transmission control elements

This section reviews open models for high-voltage direct current (HVDC) lines, grid-scale batteries, and other transmission control elements. These models play a critical role in modern power systems by enabling efficient long-distance transmission, energy storage, and improved grid stability.

1) **HVDC lines:** Both average and detailed models are available for the thyristor-based HVDC transmission system [32], [33]. These MATLAB-based models illustrate the steady-state and transient performance of a 12-pulse, 1000 MW (500 kV, 2 kA) HVDC transmission system operating at 50/60 Hz.
2) **Battery energy storage system:** In the PSCAD simulator, a general description of the functionality of the controllers and the battery system is provided and simulation results are discussed [34]. In addition, it provides a general description of the entire system and the functionality of each module. It explains how the system works and what functionality can be expected from the single-phase battery system [35]. In [36], a PV system connected to the grid with a battery system for storing energy when the PV output exceeds load consumption is discussed. The Superconducting Magnetic Energy Storage (SMES) object is also available in PSCAD simulator [37].
3) **General Voltage Source Control (VSC) simulation:** The PSCAD simulator offers a model featuring two 100 MVA PV systems [38], which can operate in either grid-following or grid-forming modes. The model also includes a simple power system connecting the PV systems, where faults can be applied.

VI. Open Computational Tools for Power Grid Simulation

Proprietary software like MATLAB, PSS/E, and PowerWorld is commonly used in academic studies for power system analysis. However, these tools impose significant limitations for AI training. Their closed algorithms and restrictive licensing hinder reproducibility and accessibility, creating barriers for collaboration and customization. Additionally, the lack

TABLE I
MODELS FOR INVERTER-BASED RESOURCES

Category	Model Type	Model Description
Type III Wind Turbine/Farm	Average	A 9 MW wind farm consisting of six 1.5 MW wind turbines connected to a 25 kV distribution system exports power to a 120 kV grid through a 30 km feeder. Wind turbines using a doubly-fed induction generator (DFIG) consist of a wound rotor induction generator and an AC/DC/AC IGBT-based PWM converter modeled by voltage sources. The stator winding is connected directly to the 60 Hz grid while the rotor is fed at variable frequency through the AC/DC/AC converter.
Type III Wind Turbine/Farm	Average and Detailed	The electrical components contain: rotor-side converter and controller, grid-side converter and controllers, DC-link system including chopper, low pass filter, crowbar protection; A scaling component is presented that can model a number of wind turbine units as an aggregated model to simulate one unit or a wind farm of several units.
Type IV Wind Turbine/Farm	Average	A 10 MW wind farm consisting of five 2 MW wind turbines connected to a 25 kV distribution system exports power to a 120 kV grid through a 30 km, 25 kV feeder. It consists of a synchronous generator connected to a diode rectifier, a DC-DC IGBT-based PWM boost converter and a DC/AC IGBT-based PWM converter modeled by voltage sources.
Type IV Wind Turbine/Farm	Average and Detailed	The mechanical components such as pitch controller and wind turbine are modeled and described. The electrical components contain: machine-side converter and controller, grid-side converter and controllers, DC-link system including chopper, low pass filter. A scaling component is presented that can model a number of wind turbine units as an aggregated model to simulate one unit or a wind farm of several units.
PV Inverter/Farm	Average	A 100-kW PV array is connected to a 25 kV grid via a DC-DC boost converter and a three-phase three-level Voltage Source Converter (VSC). Maximum Power Point Tracking (MPPT) is implemented in the boost converter by means of a Simulink® model using the 'Perturb & Observe' technique.
PV Inverter/Farm	Average	It represents a small PV farm (400 kW) connected to a 25-kV grid using two-stage converter. The PV farm consists of four PV arrays delivering each a maximum of 100 kW at 1000 W/m2 sun irradiance. A single PV array block consist of 64 parallel strings where each string has 5 SunPower SPR-315E modules connected in series.
Solar Power Plant Controller (PPC)	N/A	It is implemented that controls the overall operations of the generation plant at the point of connection (POC). The PV array generates a maximum power of 0.25MW at the nominal irradiation of 1000W/m2 and nominal temperature of 28 degrees C. A boost converter controls the DC voltage or obtains the maximum power point tracking (MPPT). The main power electronic component i.e. DC-AC inverter controls the active and reactive powers. A scaling component is introduced to model a number of inverters as an aggregated model to simulate one unit or a solar farm of several units.

of transparency and adaptability in these platforms (e.g., inexistent support for Operating Systems other than Microsoft Windows) limits the integration of novel AI techniques and tools. These challenges highlight the need for open-source alternatives that offer greater flexibility and scalability in AI research, which we review in the following.

A. Open source simulators and capabilities

Open-source simulators have become indispensable tools in the research and development of complex systems, offering flexibility, transparency, and community-driven improvements. This section explores the capabilities of various open-source simulators, highlighting their unique features, strengths, and limitations. By leveraging these tools, researchers and practitioners can model, simulate, and analyze systems across different domains, fostering innovation and collaboration while ensuring that findings are reproducible and accessible to a broader audience. This section lists and compares common open-source power simulators in Table II and Table III, focusing on their functions (maturity), documentation, and community support.

B. Open and common formats for power system data

This section explores the open and common formats used for power system data, which play a critical role in ensuring interoperability, transparency, and accessibility in power system research and applications.

1) **Common Information Model (CIM)** [58] is an essential set of open standards for representing power system components, originally developed by the Electric Power Research Institute (EPRI) and now maintained by the International Electrotechnical Commission (IEC). Comprised of IEC 61970, 61968, and 62325, CIM supports seamless data exchange within power systems and across electricity markets. Initially designed for Energy Management Systems (EMS), CIM's scope has expanded to include asset tracking, work scheduling, and market data exchanges. CIM is based on XML, which has the drawback of being very verbose and not human-readable, and data is stored on a flat structure, which is convinient for information exchange, but no so for management.
2) **Grid Research for Good (GRG) data format** [59], [60] is a JSON specification for power system data, developed within the Grid Research for Good project [61]. The specification was developed to provide a format to contain all information required to from power flows up to unit commitment problems. The specification allowed modeling topology in detail, going from node-breaker to bus-branch topologies automatically. The project is, unfortunately, no longer maintained, and lack of community support prevented its adoption in open power systems analysis tool.
3) **Common Electric Power Transmission Model (CTM)** is a power system data and structure specification intended to capture the common elements between multiple power flow and optimal power flow tools, including Matpower [43], SIIP/Sienna [40], and PowerModels [47]. The model later expanded through projects requiring unified interfaces for multiple tools, incorporating JSON and extending its scope beyond power flow applications [62]. It provides an intuitive, extensible, language-agnostic interface, framework that defines

TABLE II
List of Open-Source Power System Simulators for Transmission Systems

Name	User Interface	Language	Simulation Functions	Use Cases
GridDyn [39]	(>_) Command Line	</> C/C++	Transmission	• Power flow simulations, • Contingency analysis, • Dynamic simulations, • Regulation analysis.
Scalable Integrated Infrastructure Planning (SIIP) [40]	(>_) Command Line 🖥 Network Visualization	julia	Transmission	• Multiple *simulation packages* provide various functions for power system modeling, simulations, and visualization, • Simulation visualization package, • Network matrices for power system modeling, such as Ybus, PTDF, and LODF, • Dynamic modeling and simulations, • Market simulations, • Production cost modeling, • Integrated resource planning.
pandapower [41]	(>_) Command Line 🖥 Network Visualization	Python	Transmission, Distribution	• Power flow simulations (including 3-phase unbalanced system), • Optimal power flow simulations, • State estimation, • Short-circuit calculation, • Topology graph searches, • Contingency analysis, • Power network visualization, • Dynamic simulations.
Matpower [42], [43]	🖥 MATLAB UI	Matlab	Transmission	• Power flow simulations, • Continuation power flow, • Extensible optimal power flow, • Unit commitment studies, • Stochastic, secure multi-interval OPF/UC.
PSAT [44]	🖥 MATLAB GUI	Matlab	Transmission	• Power flow simulations, • Continuation power flow, • Optimal power flow, • Small signal stability analysis, • FACTS models, • Wind Turbine models.
PyPSA [45]	(>_) Command Line	Python	Transmission	• Power flow simulations, • Optimal power flow (linear and security-constrained), • Optimization of total electricity/energy system least-cost investment.
PowerModels.jl [46], [47]	(>_) Command Line	julia	Transmission	• Power flow simulations, • Optimal power flow, • Optimal transmission switching, • Transmission network expansion planning. NOTE: The PowerModels toolset offers multiple packages, covering optimization problems for both transmission and distribution grids.
GridCal [48]	🖥 GUI	Python	Transmission	• Power flow simulations, • Optimal power flow, • Continuation power flow, • Short-circuit calculation, • Contingency analysis, • AC linear net transfer capacity calculation, • AC+HVDC optimal net transfer capacity calculation, • Sigma analysis (one-shot stability analysis), • Investments analysis.
ANDES [49], [50]	(>_) Command Line	Python	Transmission	• Power flow simulations, • Industry-grade second-generation renewable models (solar PV, type 3 and type 4 wind), distributed PV, and energy storage model, • Out-of-the-box PSS/E raw and dyr file support for available models, • A unique hybrid symbolic-numeric approach to modeling and simulation that enables descriptive DAE modeling and automatic numerical code generation, • A rich library of transfer functions and discontinuous components (including limiters, dead-bands, and saturation) available for prototyping models, which can be readily instantiated as multiple devices for system analysis.

TABLE III
LIST OF OPEN-SOURCE POWER SYSTEM SIMULATORS FOR DISTRIBUTION SYSTEMS AND OTHER FUNCTIONS

Name	User Interface	Language	Simulation Functions	Use Cases
OpenDSS [51]	🖥 GUI	</> **C/C++**	Distribution	• Power flow simulations (unbalanced, multi-phase), • Linear and non-linear analysis, • Stray voltage/current analysis, • Dynamic/EMT simulations, • Fault analysis, • Harmonic analysis, • Flicker analysis.
GridLAB-D [52], [53]	(>_) Command Line	</> **C/C++**	Distribution	• 3-Phase, unbalanced (meshed or radial) power systems, • Reliability analysis, • Microgrid capabilities, including machine dynamics and generator controls, • Advanced control and optimization algorithms and interfaces, • API for interfacing with third-party tools, • Retail level markets and transactive controls.
PYPOWER [54]	(>_) Command Line	Python	Power flow/ optimal power flow solver	• Power flow simulations (DC and AC), • DC and AC optimal power flow simulations. NOTE: PYPOWER is no longer actively maintained.
UnitCommitment .jl [55]	(>_) Command Line	julia	Grid Optimization	• Security-Constrained Unit Commitment (SCUC) studies, • Diverse collection of large-scale benchmark instances collected from the literature, • Julia/JuMP implementations of state-of-the-art formulations and solution methods for SCUC, including ramping formulations (i.e., ArrCon2000, MorLatRam2013, DamKucRajAta2016, and PanGua2016), piecewise-linear costs formulations (i.e., Gar1962, CarArr2006, and KnuOstWat2018), and contingency screening methods (i.e., XavQiuWanThi2019).
EGRET [56], [57]	(>_) Command Line	Python	Grid Optimization	• Library of different problem formulations and approximations, • Generic handling of data across model formulations, • Declarative model representation to support formulation development, • Economic Dispatch (optimal power flow) studies (e.g., DCOPF, ACOPF), • Unit-Commitment studies.

relationships between electric power network components, supporting the development of advanced computational methods for power systems operations and simulations, across programming languages (currently, providing API's for Python, Julia, and C++).

VII. CONCLUSIONS

In this survey paper we reviewed currently available datasets for power grids and open-source simulators. We find that availability and accessibility to power grid data, while presenting major improvements from just a few years ago, remains scarce and a challenge to be tackled.

In terms of availability, two important gaps remain. First, is the nonexistence of complete virtual power grids, that is, open or synthetic power grids with all the technical parameters you would expect to find in a real system, to conduct simulations ranging from EMT up to long-term expansion planning. This absence prevents the development of techniques that could lead to better operational and planning tools, e.g., maintenance scheduling while ensuring transient stability, as those techniques would require access to both dynamic parameters and thermal parameters, which are not available for any open system at present. Second, is the lack of simulation data with labels that would be necessary to train ML models to approximately recognize patterns that are too expensive to compute exactly, e.g., network topologies that would lead to voltage instability or IBR configurations that would lead to frequency deviations upon contingencies. Recent advances in this direction include the publication of a large optimal power flow database [2], but many more advances are required before power grid data is on par with fields that ML has previously revolutionized.

In terms of accessibility, the main gap is the difficulty in parsing and understanding power system data by non experts. Current open datasets are published on a mix of formats, ranging from plain text tables (RAW format), CSV tables, Microsoft Excel files, etc., with very little documentation on how to make sense of the data outside of power grid tools. Work in the GRG and, recently, the CTM specification have the potential to advance in closing this gap, but community support remains critical for the effort to have long-term impact.

ACKNOWLEDGMENT

This work was sponsored by the U.S. Department of Energy (DOE), Office of Electricity, Transmission Reliability and Renewable Integration (TRRI) Program. It was performed under the auspices of the DOE by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

REFERENCES

[1] Grid data program. DOE ARPA-E. [Online]. Available: https://arpa-e.energy.gov/technologies/programs/grid-data

[2] S. Lovett, M. Zgubic, S. Liguori, S. Madjiheurem, H. Tomlinson, S. Elster, C. Apps, S. Witherspoon, and L. Piloto, "Opfdata: Large-scale datasets for ac optimal power flow with topological perturbations," 2024. [Online]. Available: https://arxiv.org/abs/2406.07234

[3] A. Venzke, D. K. Molzahn, and S. Chatzivasileiadis, "Efficient creation of datasets for data-driven power system applications," *Electric Power Systems Research*, vol. 190, p. 106614, 2021.

[4] Infotecnica (in spanish). Coordinador Electrico Nacional (CEN, Chile). [Online]. Available: https://infotecnica.coordinador.cl/

[5] Sistema de información de parámetros técnicos PARATEC (in spanish). Centro Nacional de Despacho (CND, Colombia). [Online]. Available: https://paratec.xm.com.co/paratec/SitePages/Default.aspx

[6] (2024, August) Open energy modeling. [Online]. Available: https://wiki.openmod-initiative.org/wiki/Main_Page

[7] F. Wiese, I. Schlecht, W.-D. Bunke, C. Gerbaulet, L. Hirth, M. Jahn, F. Kunz, C. Lorenz, J. Mühlenpfordt, J. Reimann, and W.-P. Schill, "Open power system data – frictionless data for electricity system modelling," *Applied Energy*, vol. 236, pp. 401–409, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261918318130

[8] (2024, August) Open power system data. [Online]. Available: https://open-power-system-data.org/

[9] H. Yuan, R. S. Biswas, J. Tan, and Y. Zhang, "Developing a reduced 240-bus wecc dynamic model for frequency response study of high renewable integration," in *2020 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, 2020, pp. 1–5.

[10] Test case repository for high renewable study. National Renewable Energy Laboratory (NREL). [Online]. Available: https://www.nrel.gov/grid/test-case-repository.html

[11] L. Zhu, S. You, H. Yin, Y. Zhao, F. Li, W. Yao, C. O'Reilley, W. Yu, C. Zeng, X. Deng, Y. Zhao, Y. Cui, Y. Zhang, and Y. Liu, "FNET/GridEye: A tool for situational awareness of large power interconnetion grids," in *2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, 2020, pp. 379–383.

[12] S. Kunkolienkar, F. Safdarian, J. Snodgrass, A. Birchfield, and T. Overbye, "A description of the texas a&m university electric grid test case repository for power system studies," in *2024 IEEE Texas Power and Energy Conference (TPEC)*, 2024, pp. 1–6.

[13] Electric grid test case repository. Texas A&M University. [Online]. Available: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/

[14] DR POWER. [Online]. Available: https://egriddata.org/

[15] Power Grid Lib - Optimal Power Flow. [Online]. Available: https://github.com/power-grid-lib/pglib-opf

[16] IEEE PES PSACE Committee Test Case Coordination Working Group. IEEE PES PSACE. [Online]. Available: https://site.ieee.org/pes-tccwg/links-to-test-cases/

[17] Opengridmap. Russ Garret. [Online]. Available: https://paratec.xm.com.co/paratec/SitePages/Default.aspx

[18] W. Medjroubi, U. P. Müller, M. Scharf, C. Matke, and D. Kleinhans, "Open data in power grid modelling: New approaches towards transparent grid models," *Energy Reports*, vol. 3, pp. 14–21, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484716300877

[19] Federal Energy Regulatory Commission, "2022 Assessment of Demand Response and Advanced Metering." [Online]. Available: https://www.ferc.gov/media/2022-assessment-demand-response-and-advanced-metering

[20] U.S. Energy Information Administration, "Annual Electric Power Industry Report, form eia-861 detailed data files 2013-2022." [Online]. Available: https://www.eia.gov/electricity/data/eia861/

[21] U. E. I. A. (EIA). (2022) California consumers respond to appeals for electricity conservation during heat wave. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=54039

[22] (2023, July) Ercot ancillary service data. Electric Reliability Council of Texas (ERCOT). [Online]. Available: https://www.publicportal.ercot.com/csp

[23] (2024, August) Demand response across the continental US for 2006. Open Energy Data Initiative (OEDI). [Online]. Available: https://data.openei.org/submissions/180

[24] Federal energy management program. US Department of Energy. [Online]. Available: https://www.energy.gov/femp/federal-energy-management-program

[25] Smartmeter energy consumption data in london households. UK Power Networks. [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[26] Pecan street dataset. [Online]. Available: https://www.pecanstreet.org/

[27] J. R. Vázquez-Canteli, J. Kämpf, G. Henze, and Z. Nagy, "Citylearn v1.0: An openai gym environment for demand response with deep reinforcement learning," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys '19, 2019, p. 356–357.

[28] CityLearn. [Online]. Available: https://www.citylearn.net/

[29] NSRDB: national solar radiation database. National Renewable Energy Laboratory. [Online]. Available: https://nsrdb.nrel.gov/

[30] The power project. National Aeronautics and Space Administration (NASA). [Online]. Available: https://power.larc.nasa.gov/a

[31] Total weather data. Visual Crossing Weather. [Online]. Available: https://www.visualcrossing.com/weather-data

[32] Thyristor-based hvdc transmission system average model. MathWorks. [Online]. Available: https://www.mathworks.com/help/sps/ug/thyristor-based-hvdc-transmission-system-average-model.html

[33] Thyristor-based hvdc transmission system. MathWorks. [Online]. Available: https://www.mathworks.com/help/sps/ug/thyristor-based-hvdc-transmission-system-detailed-model.html

[34] Three-phase battery system - a generic example. PSCAD. [Online]. Available: https://www.pscad.com/knowledge-base/article/462

[35] Single-phase battery system - a generic example. PSCAD. [Online]. Available: https://www.pscad.com/knowledge-base/article/434

[36] Photovoltaic-battery system. PSCAD. [Online]. Available: https://www.pscad.com/knowledge-base/article/471

[37] Superconducting magnetic energy storage (smes). PSCAD. [Online]. Available: https://www.pscad.com/knowledge-base/article/264

[38] Grid forming inverter models. PSCAD. [Online]. Available: https://www.pscad.com/knowledge-base/article/894

[39] P. Top. GridDyn simulator. Lawrence Livermore National Laboratory. [Online]. Available: https://github.com/LLNL/GridDyn

[40] Open source tools for scientific energy systems analysis. National Renewable Energy Laboratory. [Online]. Available: https://github.com/nrel-sienna

[41] pandapower. [Online]. Available: https://www.pandapower.org/

[42] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, 2011.

[43] Matpower: Free, open-source tools for electric power system simulation and optimization. [Online]. Available: https://matpower.org/

[44] Psat: The power system analysis toolbox. [Online]. Available: http://faraday1.ucd.ie/psat.html

[45] Pypsa: Python for power system analysis. [Online]. Available: https://pypsa.org/

[46] 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)*, 2018, pp. 1–8.

[47] PowerModels.jl. Los Alamos National Laboratory. [Online]. Available: https://github.com/lanl-ansi/PowerModels.jl

[48] Gridcal. [Online]. Available: https://github.com/SanPen/GridCal

[49] H. Cui, F. Li, and K. Tomsovic, "Hybrid symbolic-numeric framework for power system modeling and analysis," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 1373–1384, 2021.

[50] ANDES: Python software for symbolic power system modeling and numerical analysis. Center for Ultra-Wide-Area Resilient Electric Energy Transmission Networks (CURENT). [Online]. Available: https://docs.andes.app/en/latest/getting_started/index.html

[51] Introduction to OpenDSS. Electric Power Research Institute. [Online]. Available: https://opendss.epri.com/opendss_documentation.html

[52] D. P. Chassin, K. Schneider, and C. Gerkensmeyer, "GridLAB-D: An open-source power systems modeling and simulation environment," in *2008 IEEE/PES Transmission and Distribution Conference and Exposition*, 2008, pp. 1–5.

[53] GridLAB-D. Pacific Northwest National Laboratory. [Online]. Available: https://www.gridlabd.org/started.stm

[54] PYPOWER. [Online]. Available: https://pypi.org/project/PYPOWER/

[55] UnitCommitment.jl. Argonne National Laboratory. [Online]. Available: https://github.com/ANL-CEEESA/UnitCommitment.jl

[56] B. Knueven, J. Ostrowski, and J.-P. Watson, "On mixed-integer programming formulations for the unit commitment problem," *INFORMS Journal on Computing*, vol. 32, no. 4, pp. 857–876, 2020.

[57] Egret. [Online]. Available: https://github.com/grid-parity-exchange/Egret

[58] (2022, April) Common information model primer: Eighth edition. Electric Power Research Institute. [Online]. Available: https://www.epri.com/research/products/000000003002006001

[59] (2024, August) Grg data format. [Online]. Available: https://readthedocs.org/projects/grg-grgdata/downloads/pdf/stable/

[60] R. Bent, C. Coffrin, F. Fioretto, T. W. Mak, P. Panciatici, and P. V. Hentenryck, "GRG-Schema v4.0," 2018. [Online]. Available: https://bpb-us-w2.wpmucdn.com/sites.gatech.edu/dist/3/865/files/2018/08/schemav4.0.pdf

[61] P. V. Hentenryck, "High fidelity, year long power network data sets for replicable power system research," 2020. [Online]. Available: https://www.ferc.gov/sites/default/files/2020-08/W1-3_Hentenryck.pdf

[62] (2024, August) Common electric power transmission system model. Lawrence Livermore National Laboratory. [Online]. Available: https://github.com/LLNL/ctm