Generation of Power Network Operating Scenarios for an AI-friendly Digital Environment

José Paulos, Pedro R. Silva, Ricardo J. Bessa
Center for Power and Energy Systems
INESC TEC
Porto, Portugal
{jose.paulos, pedro.r.silva, ricardo.j.bessa}@inesctec.pt

Antoine Marot, Jerome Dejaegher, Benjamin Donnot
R&D and AI Lab
Réseau de Transport d'Électricité (RTE)
Paris, France
{antoine.marot, jerome.dejaegher, benjamin.donnot}@rte-france.com

***Abstract*—With the growing need for AI-driven solutions in power grid management, this work addresses the challenge of creating realistic synthetic operating scenarios essential for developing, testing, and validating AI-based decision-making systems. It uses spatial-temporal noise functions, predefined patterns, and optimal power flow to model renewable energy and conventional power plant generation, load, and losses. Quantitative and visual key performance indicators are proposed to evaluate the quality of the generated operating scenarios, and the validation highlights the framework's ability to emulate diverse and practical operating scenarios, bridging gaps in AI-driven power system research and real-world applications.**

***Index Terms*—Power network, artificial intelligence, synthetic data, operating scenarios, open-source**

I. Introduction

Modern artificial intelligence (AI) techniques, such as reinforcement learning (RL), are emerging as fast decision-aid tools for real-time and predictive power network control [1]. Applying AI to such systems requires open-source digital environments that emulate realistic physical system operations and human decision-making, enabling effective AI development, testing, and deployment while directly assessing decision quality. One example is the Grid2Op open-source environment, developed by RTE, models a wide range of power system problems, particularly congestion management, and supports the development and evaluation of grid controllers [2].Through different L2RPN (Learning to Run a Power Network) competitions [2], calibrated virtual environments have been instantiated for testing over robustness to adversarial attacks or adaptability to increasing renewable energy share. Similarly, the Grid Optimization (GO) competition [3] explored real-time, day-ahead, and week-ahead market applications but was limited to a few scenarios. Other AI-friendly environments include Flatland [4] for railway scheduling and BlueSky [5] for realistic air traffic simulation.

The research leading to this work is part of the AI4REALNET (*AI for REAL-world NETwork operation*) project, which received funding from European Union's Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527, and from the Swiss State Secretariat for Education, Research and Innovation (SERI). This project is funded by the European Union and SERI. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union and SERI. Neither the European Union nor the granting authority can be held responsible for them.

A key building block of these environments is the capacity to generate realistic synthetic time series data from publicly available datasets, preserving the spatial-temporal dependencies of the original data. Synthetic data is critical for training and validating RL-based agents across diverse grid operating scenarios, even when real data is available. Traditional methods, such as ARIMA [6] and Copula theory [7], have been applied to model spatial-temporal dependencies. Recent advances in generative AI, including generative adversarial networks (GANs) [8] and diffusion models [9], enable the generation of energy time series, such as photovoltaic and load data, at various temporal resolutions and grid levels. Furthermore, other works [10], [11] focus on synthetic grid generation using OpenStreetMap data and apply two distinct approaches: a) an electrical algorithm combined with the domain expertise to assign consumers to nodes and define parameters for lines and transformers, and b) combining a Bayesian Hierarchical Model with Markov Chain Monte Carlo techniques.

However, testing AI-based power grid controllers requires building coherent operating scenarios that extend beyond spatial-temporal time series generation. These scenarios must incorporate additional elements, such as conventional generation dispatch, post-processing for simulated losses, and network limits that address challenges from the energy transition (e.g., higher integration of renewable energy sources). This paper introduces advancements in generating realistic synthetic operational scenarios using the Grid2Op environment [12], a scenario-based simulator designed for training and evaluating AI recommendation agents. The methodology enhances the approach in [13] by improving multivariate time series sampling and adding a conventional generation dispatch function. Related works include [14], which manually constructs a synthetic transmission grid for ERCOT 2030 with RES investments, and [15], which combines statistical modeling and optimization to generate realistic load, renewable production, and generation offers. Compared to these works, the main innovation from the present paper is the generation of operating scenarios specifically tailored to AI applications, particularly for gymnasiums where RL agents are trained and validated.

The paper is organized as follows: Section II details the scenario generation methodology. Section III introduces key performance indicators (KPIs) for evaluating scenario quality, which are used in Section IV to validate the generated time series. Section V presents the conclusions.

II. Methodology for Scenarios Generation

This section presents a comprehensive framework, available as the open-source tool *ChroniX2Grid* [16], for generating synthetic data that emulate realistic grid scenarios, including renewable energy sources (solar and wind), electrical consumption (load), network losses, and economic dispatch for conventional power plants. Pattern-based methods with spatial-temporal noise are used for solar, wind, and load data, while losses are generated using a pattern-based approach and further refined with AC load flow after solving the dispatch problem. Economic dispatch is determined through an Optimal Power Flow (OPF) solver and depends on the prior generation of solar, wind, load, and losses to ensure a fully characterized system.

A. Pattern-based with Spatial-temporal Correlated Noise

This method applies a generated noise to a reference pattern, aiming to host deviations (noise) in the original synthetic diagram. This is applied independently of the category of data to be generated: $pattern^{solar}$, $pattern^{wind}$, and $pattern^{load}$). In this context, since the goal is to generate data for a network with geographic distribution and with data generation over a certain time horizon, two types of noise are applied hence, giving us a spatial-temporal noise function – $f^{category}_{t(x,y)}$ – depending on the category of data to be generated. This function is based on the following components: i) a three-dimensional mesh (x,y,t) with independent noise points, and ii) Spatial and temporal interpolation of this noise at a given geographical location and timestep.

In short, the spatial-temporal noise is generated geographically and in a coarse timeline and then interpolated for the final dataset's requested resolution dt. The noise function generation process is divided into three main stages, detailed in the following subsections.

1) Three-dimensional mesh conception: A two-dimensional reference mesh is built for each coarse time step to generate noise functions. Given the geographical distribution of the network, the initial dimensions and granularity follow the user-provided parameter data and can then be adjusted or updated. This adjustment procedure aims to increase the default mesh size to cover all network nodes. This overall spatial architecture of the mesh considers two main parameters: i) the total length of the mesh (Lx, Ly), ii) the granularity of the mesh (dx_{corr}, dy_{corr}), representing the distance in which the spatial phenomena are independent. Since it should also consider temporal noise, this mesh is then made recurrent in temporal layers, at a given coarse time resolution, depending on the generation category ($category_{corr}$). The objective of having a checkered mesh (with virtual rows and columns) is to guarantee that every network node can directly relate to the 4-nearest mesh neighborhood. Since any provided network could have an infinite combination of geographical coordinates, this formulation can adjust the mesh's size accordingly to cover all network nodes. To achieve this, the process identifies the largest network coordinates and determines the number of rows (x_{plus}) and/or columns (y_{plus}) that must be added to the original mesh dimensions, as described in Eq. 1.

$$x_{plus} = int\left(\frac{x}{dx_{corr}}\right)+1, \quad y_{plus} = int\left(\frac{y}{dy_{corr}}\right)+1 \quad (1)$$

The final mesh structure should virtually look like the example illustrated in Fig. 1. To each mesh node, a Gaussian noise is automatically generated following a $N(0,1)$ distribution, hence creating the $NoiseNode_{i,t}$.

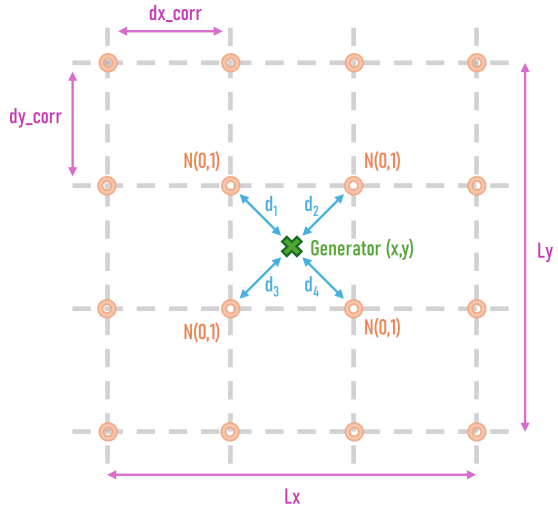


Fig. 1. Example coarse noise mesh for a single coarse timestep

2) Auto-correlated mesh noise: After establishing the complete mesh with node-independent Gaussian noise, the next stage aims to produce noise based on the distance to the 4-nearest neighbors for each generator node. To prevent very long computational execution times in scenarios with a large number of generators, this methodology takes advantage of the spatial and time coarse mesh resolution that is interpolated in later stages. Following this, the next step is to apply the spatial noise to each generator node. As seen in Fig. 1, four distances are calculated for the respective four-nearest neighbors ($ng, i, \in [1,4]$) using Eq. 2.

$$dist_{i,t} = \frac{1}{\sqrt{\left(x - dx_{corr}\cdot x^{ng}_{i,t}\right)^2 + \left(y - dy_{corr}\cdot y^{ng}_{i,t}\right)^2} + 1} \quad (2)$$

These four distances provide sufficient information to compute the generator noise corresponding to each of the four nearest neighbors (i), as defined in Eq. 3.

$$GenNoise_t = \frac{\sum_{i=1}^{4} dist_{i,t}\cdot NodeNoise_{i,t}}{\sum_{i=1}^{4} dist_{i,t}} \quad (3)$$

3) Temporal Interpolation: The last step of this methodology is based on a classic interpolation (linear, quadratic, or cubic). This process allows refining the coarse noise timeline created with t spatial layers (based on $category_{corr}$) to the required temporal resolution. With this, the generated noise can now cover all original data measurements since the resolution matches accordingly. This final noise represents the $f_{t(x,y)}^{category}$ that is applied to the respective *category* at a later stage.

B. Reference Patterns

The generation of solar and wind energy data, load, and losses is guided by a predefined dataset (e.g., open data of generation profiles from ENTSO-E Transparency Platform) containing a generic yet representative profile of these generation patterns. These are classic daily to yearly patterns climbing up during sun hours and also peaking their values during the summer. For each category, there is a $pattern_t^{category}$, variable in t, with annual duration and hourly resolution. It is important to note that, in the case of wind, there are three datasets (or profiles) for three different time horizons (short, medium, and long) and, thus, three different noise functions. These patterns are also internally interpolated for the final dataset requested resolution dt and repeated for as many years as necessary.

C. Independent Additional Noise

Apart from the spatial-correlated noise, an additional noise n_s could also be added. In this sense, n_s is formulated using a regular Gaussian distribution following $N(0, s)$, where s represents the adjustable standard deviation of the noise. The goal is to apply random and small perturbations to the system, representing very short-term weather variability. In this case-scenario, this is only applied to wind generation.

D. Solar Energy

The solar data generation follows the methodology detailed in sections II-A and II-B and hence retrieves the respective $f_t^{solar}(x, y)$ that represents the spatial-temporal correlated noise and the $pattern_t^{solar}$ with the reference solar data pattern. The actual solar data for a given generator placed in coordinates x, y, for each time step t will be created using the formulation presented in Eq. 4.

$$solar_t(x, y) = P_{max} \cdot \mathcal{S}\left(pattern_t^{solar} \cdot \left(\beta + \sigma \cdot f_t^{solar}(x, y)\right)\right) \tag{4}$$

where P_{max} is the installed solar generator capacity; $\mathcal{S}(x)$ is a smoothing function of form $\mathcal{S}(x) = 1 - exp(-x)$, to normalize data and operating a convex distribution transformation with the objective to better fit realistic data; β (defaulted as 0.75 representing 75% efficiency) is the bias of the spatial-temporal correlated noise, works as a way of preventing an around-zero distribution, and also is a strategy of modeling the efficiency of the generator; σ (defaulted as 0.8) is the spatial-temporal noise standard-deviation that works as a weight of the noise to be applied to the pattern.

E. Wind Energy

Wind data generation also follows the same methodology and requires spatial-temporal correlated noise and reference wind data patterns. This process differs from the solar generation methodology by requiring three different parameters. So, in section II-A, instead of a single $wind_{corr}$, three values are inserted ($shortWind_{corr}$, $mediumWind_{corr}$, $longWind_{corr}$) and, thus, three spatial-temporal noise functions are, respectively, extracted. These three components represent the spectrum of wind's different behavior over time with the goal of adding several types of scaling noise in all periods. Although solar energy methodology uses a third-party static pattern file, for wind data generation, the pattern is constructed relying on a simple one-year oscillating cosine (seasonal pattern) and another constant component, as described in Eq. 5.

$$pattern_t^{wind} = weight_{const} + weigth_{oscil} \cdot pattern_{seas.,t}^{wind} \tag{5}$$

The seasonal pattern is based on a full-cycle yearly cosine associating the periods with most wind with the highest cosine values (see Eq. 6).

$$pattern_{seas.}^{wind} = cos\left(\frac{2\pi}{365 \cdot 25 \cdot 60} \cdot (t - 30 \cdot 24 \cdot 60 + \delta t)\right) \tag{6}$$

This formulation considers $\frac{2\pi}{365 \cdot 25 \cdot 60}$ as one-year full cosine cycle in minutes. It also considers December as the period with the highest wind speed, granting the need to subtract one month of minutes equivalent ($30 \cdot 24 \cdot 60$) to the cumulated simulation time (t). This new time value should be reset to the year's starting point to guarantee real day/seasonal matching. For this, it sums δt, which represents the time difference between the first simulation period and January 1st, which is a standard fixed reference for the beginning of the year. Both $weight_{const}$ and $weigth_{oscil}$ values assignment should follow regional characterization but are defaulted as 0.7 and 0.3, respectively. This roughly shows that 70% of the wind generation is somehow sustained, but 30% is associated with seasonal elements and subject to changes.

The actual wind generation follows a mathematical approach close to the one applied to the solar generation but considers the three spatial-temporal noise components – short, medium, and long – independently (see Eq. 7).

$$\begin{aligned} wind_t(x, y) = P_{max} \cdot \mathcal{S}\big(0.1 \cdot exp\big(4 \cdot pattern_t^{wind}\big) \cdot \big(\beta \\ + \sigma^{shortWind} \cdot f_t^{shortWind}(x, y) \\ + \sigma^{mediumWind} \cdot f_t^{mediumWind}(x, y) \\ + \sigma^{longWind} \cdot f_t^{longWind}(x, y)\big) \\ + n_s\big) \end{aligned} \tag{7}$$

This formulation takes into account the same structure as the one in section II-D with slight differences. The parameter β now defaults to 0.3, representing 30% efficiency and σ^{type}

now considers three types (short, medium, long), defaulted to 0.02, 0.15, and 0.15, respectively.

F. Electrical Energy Consumption

This process follows the main steps of solar generation and the methodology from sections II-A and II-B and hence retrieves the respective $f_t^{load}(x, y)$ that represents the spatial-temporal correlated noise, and the $pattern_t^{load}$ with the reference load data pattern. In addition to the usual typical pattern dataset, it also considers an external dataset with a representative weekly consumption profiling - $pattern_{week,t}^{load}$, with a classic pattern (lower during weekends, and two daily peaks in the morning and in the afternoon). The main $pattern_t^{load}$ represents the seasonal pattern and is modeled on an oscillating cosine over a period of one year, and considers a constant and an oscillating component with different weights (see Eq. 8).

$$pattern_t^{load} = weight_{const} + weigth_{oscil} \cdot pattern_{seas.,t}^{load} \quad (8)$$

The assignment of $weight_{const}$ and $weigth_{oscil}$ values assignment should follow regional characterization, but defaults as $^{5.5}/_{7}$ and $^{1.5}/_{7}$, respectively. This roughly shows that 79% of the wind generation is somehow constant, but 21% is associated with seasonal elements and subject to changes. As in section II-E, the seasonal pattern also considers a specific part of the year with the highest consumption rate (mid-February), and hence requires a shift of 45 days (or $45 \cdot 24 \cdot 60$, in minutes). The remainder of the formulation remains the same (see Eq. 9).

$$pattern_{seas.}^{load} = cos\left(\frac{2\pi}{365 \cdot 25 \cdot 60} \cdot (t - 45 \cdot 24 \cdot 60 - \delta t)\right) \quad (9)$$

The load data generation process complies with the previous overall methodologies but with a simpler approach, and considers the formulation detailed in Eq. 10.

$$load_t(x, y) = P_{max} \cdot pattern_{week,t}^{load} \cdot \left(\sigma f_t^{load}(x, y) + pattern_t^{load}\right) \quad (10)$$

G. Network Losses

In order to account for network losses, an external yearly loss pattern dataset is used. It contains a static pattern with 5-min resolution with absolute active network losses. If the period to generate is larger than one year, the pattern is repeated to cover the request generation horizon. This aims to provide a rough estimate assisting the dispatch problem solving.

H. Conventional Generation

This subsection details the data generation process of hydro, nuclear, and thermic generators that follow a classic economic dispatch process. The goal is to use the previously generated data and optimize the network's power flow to minimize generation costs while meeting the demand and adhering to the operational and other technical constraints (e.g. prioritize low-cost sources, optimize stored resources, ensure reliability, etc.). In this sense, this module uses *PyPSA* [17] library for running the dispatch and requires the following input data:

- Non-conventional generator data (previously generated solar, wind, load, and losses diagrams).
- Hydro reference pattern file representing seasonality and minimum/maximum hydraulic stocks.
- Ramp mode definition (for relaxation purposes): a) *hard*, all ramp constraints are considered; b) *medium*, thermal ramp constraints are skipped; c) *easy*, thermal and hydro ramp constraints are skipped; d) *none*, all ramp constraints are skipped.
- Generator specifications (i.e. max and min power).
- Other optimization parameters (e.g., solver, reactive compensation, etc.).

The overall process aims to generate production diagrams for all missing generators, and it is divided into three main phases:

1) Pre-dispatch: planning stage to define non-conventional generation levels and, hence, the load that needs coverage; hydro generation limits; other input data structuring.
2) Dispatch: simulation of near-real-time control and operation of the network, granting conventional generation adjustments to match the structured demand and the other power system pre-defined conditions. This main stage runs an optimization module in a row for each day, week, or month, using a provided time resolution.
3) Post-dispatch: adjustment/interpolation of measurements (if necessary); marginal costs calculation; calculation of other dispatch data and statistics (e.g. total run time, data structuring, etc.).

After computing the economic dispatch, some adjustments are applied to account for transmission losses by simulating the actual energy required to cover the ordinary demand plus the network losses. With the simulation of the network power flow, it is possible to extract a close-to-real loss value and make it accountable on top of the actual load consumption from the network. Considering the updated network losses, the production level of the slack generator is adjusted to account for these more accurate values, ensuring that the system's power balance reflects a closer-to-reality representation of its operation. It also considers a stoppage criterion to address previous constraint violations.

III. Key Performance Indicators

The goal of the KPIs is to assess the quality and relevance of these synthetic data generated for integration with other third-party power system simulations. They evaluate the data in two possible scenarios: a) Load and non-conventional generation, which make use of reference diagrams (consumption/generation) from three locations in France [16], using data from *Renewable Ninja* [18]; b) All data (full dispatch), which makes use of reference diagrams (consumption/generation) from three France locations, using data from

Réseau de Transport d'Électricité (RTE), directly included in *Chronix2Grid* reference data [19]. This approach uses these real data sources as references to compare the generated data directly. While some results are quantitative, others are purely observational/qualitative. The main analysis is processed differently, depending on the data and respective KPI.

1) Conventional generation dispatch: Simplistic analysis of the overall dispatch arrangement and the respective individual amount for each type. For that, a distribution comparison is established between the generated and reference data, which is shown in a pie chart.

2) Solar and wind power generation: Verifies if the generated data has the same aggregated distribution as the provided reference data. The synthetic solar data are also verified to check if none is present during the night. In order to compare distributions two indicators are used: Kullback-Leibler Divergence (KL Divergence) and Jensen-Shannon Divergence (JS Divergence), see Eq. 11 and Eq. 12. The KL Divergence is useful for understanding how the synthetic data captures the reference data's key characteristics, and the JS Divergence comprehends symmetric comparisons and also captures the interpretability.

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) \tag{11}$$

where P and Q represent the reference and synthetic sets, respectively. Here, a smaller value indicates that the synthetic data closely resembles the reference data. It should be noted that KL Divergence is asymmetric, meaning $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$.

$$D_{JS}(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M) \tag{12}$$

where $M = \frac{1}{2}(P + Q)$ is the average or midpoint distribution. This metric checks how similar the two distributions are by comparing them to their average distribution M. The result is bounded between 0 (identical distributions) and 1 (maximally different distributions).

For these two types of generation, the Pearson correlations between generators are also calculated to evaluate the diversity and spatial dependency structure between the time series data.

3) Electrical energy consumption: Comparison of normalized total load by each week and hour of the year, where $L_{norm} = \frac{L}{L_{max}}$. With these values, the Mean Absolute Error (MAE) is computed, and since the values are normalized, it gives a percentage similarity error estimate between the reference and synthetic series.

IV. Validation of the Scenarios

To validate the generated scenarios, the network and other specifications from the `l2rpn_idf_2023` environment of the Grid2Op is selected, which was used in the Paris Region AI Challenge for Energy Transition [20]. It includes an adapted version of the IEEE 118 grid with high renewable penetration representing the target electrical mix reached by 2035 in France; the generated data (one year) is compared with reference 2012 data. The comparison with 2012 data serves only as an example to highlight differences in the mix KPI. All other KPI primarily focus on distribution's analysis to assess the quality of the generated data. Therefore, the choice of reference year should not be impactful, as we are not comparing the same networks, and both input and reference data are adjustable. For this use-case scenario, all parameters remain set to their default values within this specific environment [16].

A. Conventional Generation Dispatch

In this specific example, the proportion of nuclear generation is below the one found in the reference data, but it shows that the synthetic mix would retrieve a different (but optimal) dispatch value when compared to the reference data, as depicted in Fig. 2. Following French energy goals for 2050 [21], where the goal is to surpass 50% of nuclear, this generation structures the distribution of global generation for roughly 50% of nuclear generation. The final dispatch data for the provided period is shown in Fig. 3.

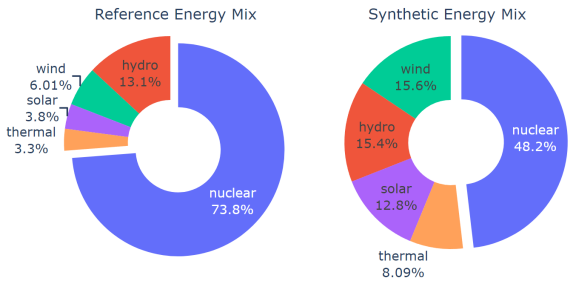


Fig. 2. Reference vs. Synthetic Energy Mix

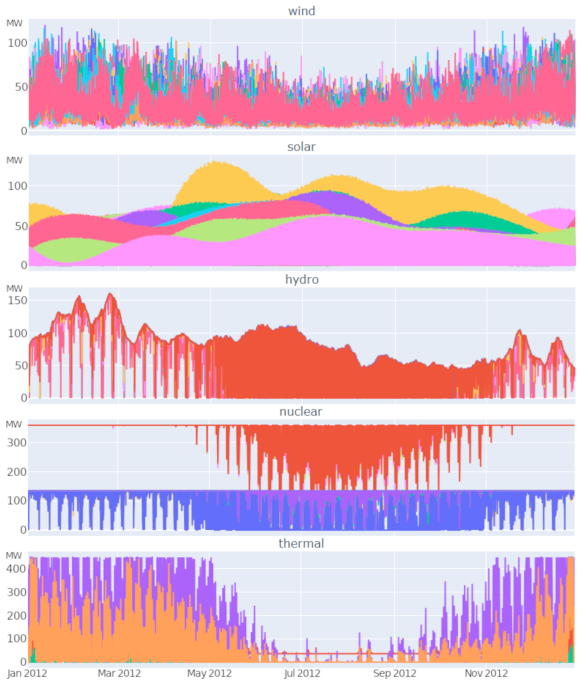


Fig. 3. Generators synthetic data for one year (one color per gen.)

For a closer look, see Fig. 4 for a random one week snapshot of the generated data for aggregated electric energy consumption, solar and wind generators.

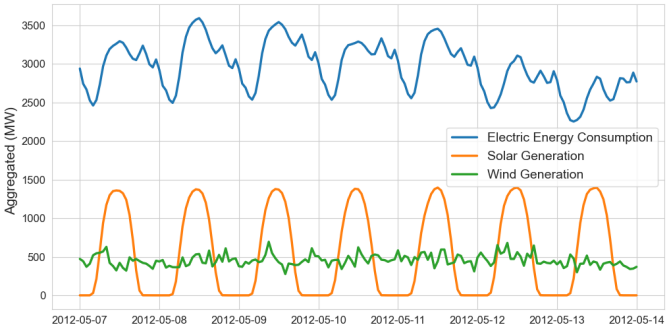


Fig. 4. Synthetic Electric Energy Consumption vs. Solar and Wind Generation - One Week Snapshot

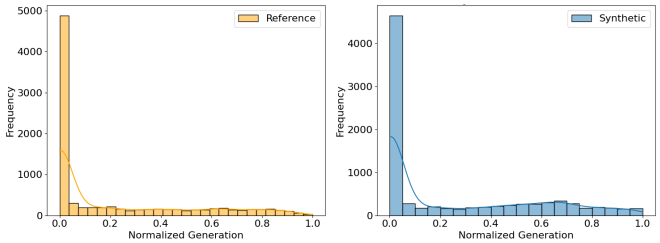
B. Solar and Wind Power Generation

For this specific use case, Fig. 5 shows that the overall distributions are similar (the y-axis is given in frequency percentage). As for the wind generators, Fig. 6 shows that the synthetic data is smoother, more symmetrical, and less variable (the y-axis is given in frequency percentage).

The results in Table IV-B indicate the degree of similarity between synthetic and reference distributions for solar and wind data. For the solar dataset, both KL and JS Divergences are low, suggesting a strong similarity between the synthetic and reference distributions. When zero values are excluded from the solar datasets, both KL and JS divergences increase, indicating that the synthetic data deviates slightly from the reference distribution in non-zero regions but conserves consistency. Both Divergences are slightly higher for the wind dataset, implying that while consistent, the synthetic data is less aligned with the reference distribution.

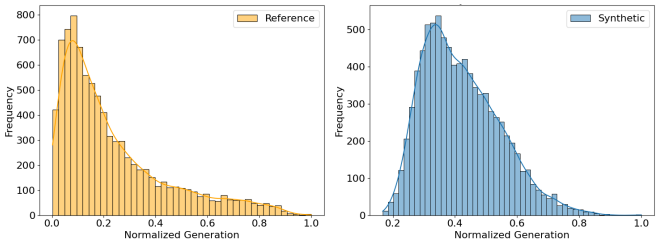
TABLE I
DIVERGENCES FOR SOLAR AND WIND POWER DATA DISTRIBUTIONS

Dataset	KL Divergence	JS Divergence
Solar	0.027	0.085
Solar w/o zeros	0.204	0.225
Wind	0.683	0.328



(a) (b)

Fig. 5. Distribution of solar production for reference (a) and synthetic (b) data



(a) (b)

Fig. 6. Distribution of wind power generation for reference (a) and synthetic (b) data

As for the Pearson's correlation matrices, the results are shown in Fig. 7 and Fig. 8, for solar and wind generation, respectively. Most solar generators show very high correlation coefficients (0.8 to 1.0), suggesting a strong similarity in their output patterns. This is consistent with overall real-world and reference solar data, where generators in similar geographic locations or under the same environmental conditions tend to have highly correlated outputs. It should also be noted that *SGen13* shows a decrease in correlation (e.g., 0.4-0.7 with some generators), representing a different profile, or even simulating environmental diversity such as shading effects, microclimate variations, or differences in panel orientation.

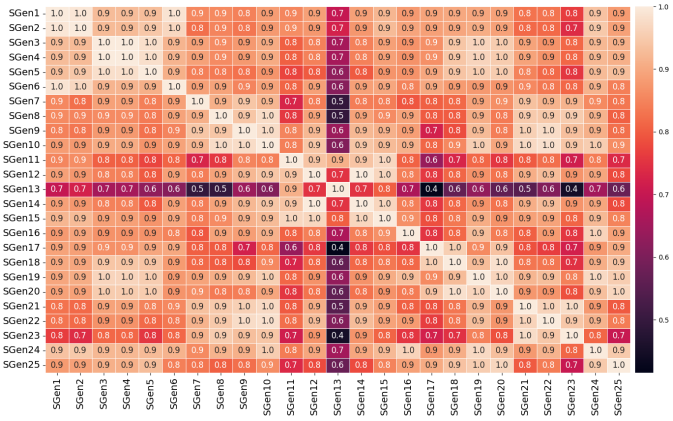


Fig. 7. Pearson correlation for solar power generators synthetic data

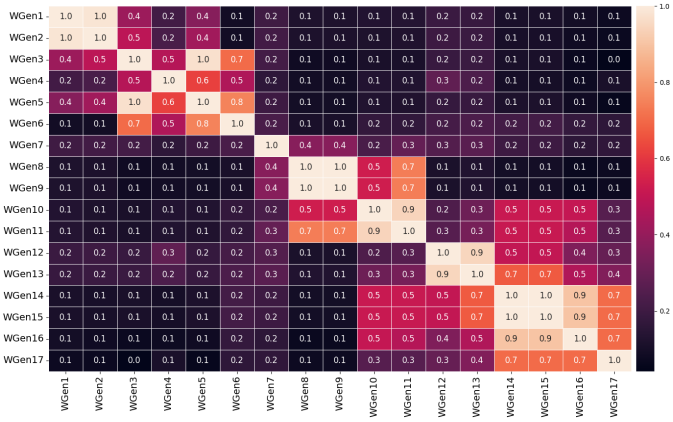
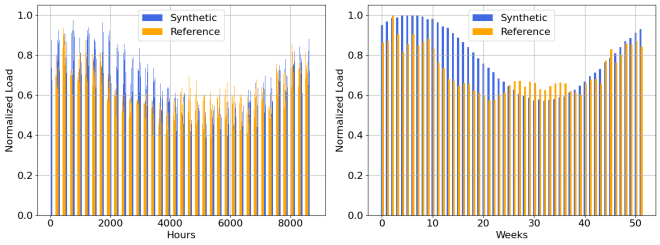


Fig. 8. Pearson correlation for wind power generators synthetic data

Unlike the solar synthetic data, the wind generator correlation matrix (Fig. 8) shows a broader range of values (from 0 to 0.9), indicating greater diversity in the generated profiles. This reflects expected behavior, as wind generation is often influenced by localized wind patterns, terrain, and turbulence, leading to less uniformity compared to solar generators. Certain generator clusters (e.g., *WGen5–WGen6* and *WGen13–WGen14*) exhibit moderate to high correlations, suggesting that these generators share similar environmental conditions and/or are located within the same wind farm.

C. Electrical Energy Consumption

As shown in Fig. 9, the reference data show a noisier and less symmetric U-shaped distribution, with a sharper drop and smaller mid-year values. The synthetic data display a smoother and more symmetric U-shaped curve, with a gradual decline to the mid-year trough and consistent recovery towards the year's end. The reference data have more pronounced weekly fluctuations, indicating higher variability, while the synthetic data appear more averaged. Both capture a similar overall trend, but the synthetic data offers a more idealized and uniform representation of the seasonal pattern.



(a) (b)

Fig. 9. Electric energy consumption by hour (a) and week (b) of the year

For these specific scenarios, the results are quite positive with percentage equivalent $MAE_{hours} = 8.19\%$ and $MAE_{weeks} = 7.69\%$.

V. Conclusions

In recent years, the release of new datasets, benchmarks, competitions, and AI-friendly digital environments has significantly advanced AI research and development in power system operations. These initiatives are already catalyzing progress in applications such as forecasting and grid management, as exemplified by the L2RPN and GO competitions. This paper introduced a methodology for generating realistic synthetic operating scenarios using the Grid2Op environment, addressing challenges such as renewable integration, economic dispatch, and loss modeling. However, there remains a need for more diverse datasets and benchmarks, including both large-scale, open-source synthetic datasets and those incorporating real-world data. These resources support initiatives like the European Commission's emerging AI Testing and Experimentation Facilities.

Future work will compare pattern-based methods with recent advancements in generative AI for time series, evaluating key properties such as accuracy and computational efficiency, which are essential for AI-compatible digital environments.

References

[1] A. Marot, A. Rozier, M. Dussartre, L. Crochepierre, and B. Donnot, "Towards an AI assistant for human grid operators," in *Hyb. Human Art. Intel.*, 2022.

[2] A. Marot, B. Donnot, G. Dulac-Arnold, A. Kelly, A. O'Sullivan, J. Viebahn, and et al., "Learning to run a power network challenge: a retrospective analysis," in *NeurIPS Comp. and Demo. Track*, vol. 133, 2020, pp. 112–132.

[3] F. Safdarian, J. Snodgrass, J. H. Yeo, A. Birchfield, C. Coffrin, and C. Demarco, "Grid optimization competition on synthetic and industrial power systems," in *NAPS 2022*, Salt Lake City, UT, USA, Oct. 2022.

[4] F. Laurent, M. Schneider, C. Scheller, J. Watson, J. Li, Z. Chen, and et al., "Flatland competition 2020: MAPF and MARL for efficient train coordination on a grid world," in *NeurIPS Comp. and Demo. Track*, vol. 133, 2020, pp. 275–301.

[5] J. Groot, G. Leto, S. Vlaskin, and A. Moec, "BlueSky-Gym: Reinforcement learning environments for air traffic applications," in *SESAR Innovation Days*, 2024.

[6] B. Klöckl, "Multivariate time series models applied to the assessment of energy storage in power systems," in *PMAPS 2008*, Rincón, Puerto Rico, May 2008.

[7] R. Becker, "Generation of time-coupled wind power infeed scenarios using pair-copula construction," *IEEE Trans. on Sust. Energy*, vol. 9, no. 3, pp. 1298–1306, Jul. 2018.

[8] A. Pinceti, L. Sankar, and O. Kosut, "Generation of synthetic multi-resolution time series load data," *IET Smart Grid*, vol. 6, no. 5, pp. 492–502, Oct. 2023.

[9] N. Lin, P. Palensky, and P. P. Vergara, "EnergyDiff: Universal time-series energy data generation using diffusion models," *arXiv:2407.13538*, pp. 1–10, 2024.

[10] C. S. Dande, L. Mattorolo, J. da Silva Andre, L. Lavecchia, N. Efkarpidis, and D. Toffanin, "Synthetic grid generator: Synthesizing large-scale power distribution grids using open street map," *arXiv:2408.04923*, pp. 1–8, 2024.

[11] H. O. Caetano, L. D. N., M. Fogliatto, C. D. Maciel, V. P. Ribeiro, and J. A. P. Balestieri, "Bayesian hierarchical model to create synthetic power distribution systems," *Elect. Pow. Sys. Res.*, vol. 235, p. 110706, Oct. 2024.

[12] B. Donnot, "Grid2Op - A testbed platform to model sequential decision making in power systems." 2020, accessed: Jan. 2025. [Online]. Available: https://GitHub.com/Grid2Op/grid2op

[13] B. Donnot, "Deep learning methods for predicting flows in power grids: Novel architectures and algorithms," Ph.D. dissertation, Université Paris Saclay, 2019.

[14] F. Safdarian, S. Kunkolienkar, J. Snodgrass, A. Birchfield, and T. J. Overbye, "Creating a portfolio of large-scale, high-quality synthetic grids: A case study," in *IEEE KPEC 2024*, USA, Apr. 2024.

[15] M. Chatzos, M. Tanneau, and P. V. Hentenryck, "Data-driven time series reconstruction for modern power systems research," *Elect. Pow. Sys. Res.*, vol. 212, p. 108589, Nov. 2022.

[16] B. Donnot, "ChroniX2Grid: A framework for generating synthetic time series data for power grid scenarios," 2023, accessed: Jan. 2025. [Online]. Available: https://github.com/BDonnot/ChroniX2Grid

[17] T. Brown, J. Hörsch, and D. Schlachtberger, "PyPSA: Python for power system analysis," *J. of Open Res. Sof.*, vol. 6, no. 1, p. 4, 2018.

[18] Renewable Ninja Team, "Renewable Ninja: A tool for synthetic renewable power generation data," 2020, accessed: January 2025. [Online]. Available: https://www.renewables.ninja/

[19] RTE France, "RTE éCO2mix Data," https://www.rte-france.com/eco2mix/telecharger-les-indicateurs, 2025, accessed: Jan. 2025.

[20] A. Marot, L. Crochepierre, K. Chaouache, B. Donnot, A. Pavao, and I. Guyon, "Paris region AI challenge for energy transition. Low-carbon grid operations," RTE, Tech. Rep., Apr. 2023.

[21] Sénat Français, "Projet de loi de finances pour 2023: version consolidée," https://www.senat.fr/leg/tas22-040.pdf, 2023, accessed: Jan. 2025.