OALS: An Open source AI-Datacenter Load Simulator for Grid Impact Analysis

Amulya Sreejith, Balasubramaniam Natarajan
Electrical and Computer Engineering
Kansas State University
Manhattan, USA
amulya1@ksu.edu

***Abstract*—Artificial intelligence (AI) is a transformative technology that is having an increasing impact on our lives. The growing use of AI is driving the need for data centers around the world. Data centers supporting AI applications require a considerable amount of energy with some projections indicating that AI's energy needs could account for as much as 21% of all electricity usage by 2030. Therefore, characterizing these data center loads is critical for both generation planning and system operations. This paper presents a generic open source load model for AI datacenters that can be used for grid-level analysis and simulation. The proposed framework characterizes AI datacenter behavior through a modular representation integrating the temporal dynamics of IT and stochastic server workloads. By capturing correlations between IT load, facility power, and Power Usage Effectiveness (PUE) under diverse operational scenarios, the model can be integrated with both steady-state and transient grid studies. The model is validated using real and synthetic workload traces collected from GPU-based inference clusters, demonstrating strong agreement with measured power profiles and sub-minute load fluctuations. Grid impact studies on a 39 bus New England test system further illustrates its application in assessing impacts of datacenters on frequency dynamics. The model aims to provide researchers and system planners with a transparent, extensible tool that can be used to evaluate the interaction between AI datacenters and evolving electric grids.**

***Index Terms*—Large Language Models, Datacenter Power, Electrical Grid, Load Modeling, Open Source, Energy Optimization, Demand Response.**

I. Introduction

AI use is expanding rapidly across nearly every sector transforming how we work, communicate and make decisions. This surge in AI adoption has fueled an unprecedented demand for various types of data centers that process and store massive volumes of data. Unfortunately, these data centers are energy intensive, consuming vast amounts of electricity for computation, cooling etc. placing significant stress on the power grid. Sustainable and efficient integration of any large load like a data center with the power grid requires comprehensive characterization of the load behavior.

AI data centers workloads can be dynamically distributed across multiple geographically separated sites to balance performance, cost and energy efficiency. Distributed operation means that computing demand can shift spatially across regions and temporally throughout a day and week (based on user activity, energy prices and network latency). These temporal and spatial profiles of datacenter electricity demand are highly volatile, burst-prone, and stochastic. Unlike traditional IT workloads characterized by relatively stationary consumption, AI-driven compute, memory, and cooling loads exhibit strong correlations with batch scheduling, GPU utilization cycles, and network congestion patterns. As a result, the aggregate load on the grid becomes highly variable and stochastic, making it challenging to predict and manage power demands in real time.

Existing tools for grid analysis typically rely on static or generic load representations without incorporating the dynamic signatures of AI workloads. Consequently, there is a need for representative, scalable, and open-source AI-datacenter load models to better assess system-level impacts such as frequency deviations, voltage transients, and congestion interactions under realistic AI operation cycles. To address this gap, we present OALS- Open-source AI-Datacenter Load Simulator, a framework designed to emulate high-resolution power demand patterns driven by AI computation traces. The simulator integrates synthetic and empirical workload characterizations with power system testbeds, enabling reproducible analysis of grid responses to AI-induced load dynamics.

OALS supports hybrid modeling through parametric and data-driven modules, allowing flexibility to capture time-dependent GPU cycles, cooling feedback loops, and workload-switching regimes. By providing an open-source foundation, OALS promotes transparent benchmarking, collaborative model refinement, and cross-domain integration between power systems and datacenter researchers. The tool thus acts as a unifying platform bridging AI workload telemetry with electrical network simulation domains.

A. Related Work

Recent research in AI-driven datacenter power systems emphasizes predictive and real-time power management to enhance efficiency and reduce operational and environmental costs [1]. Studies by Wang et al. highlight the role of datacenters in grid-aware load balancing, turning variable demand into a controllable resource for system stability and economic benefit [2]. Machine learning based forecasting and workload allocation methods have been used for internet datacenter adaptability [3] but these models are not representative of

979-8-3315-8138-1

GPU-dominated large AI datacenters. Recent sustainability-focused models address carbon-aware operation and multi-site optimization [4]. Testbeds that use hardware like Converter-based emulators [5], [6] and hierarchical VPP/control architectures [7]. have been used to reproduce sag and load-variation responses in the grid. Despite these advances, most existing models emphasize energy efficiency or economic coordination at the IT or facility level, while limited attention has been given to the dynamic grid-level implications of AI workload variability datacenter integration [8]. For security purposes, data center operators do not share detailed hardware emulated models. In order to advance research and help develop solutions to efficiently manage the grid, it is important to create open source dynamic models that can be used to analyze grid impacts as well as control schemes for stable grid operations. Table.I shows the properties of existing methods and the proposed OALS including existence of an AI GPU model, implementation of sub-second load dynamics, IT workload to power consumption, and open-source availability.

TABLE I: Comparative analysis

Model category	AI GPU model	Sub-sec dynamics	IT to Power	Open-source
Load forecasts [1], [2]	x	x	✓	x
ML-based workload [3]	x	x	✓	x
Sustainability models [4]	x	x	✓	x
Hardware/ Emulator [5], [6]	✓	✓	✓	x
OALS	✓	✓	✓	✓

B. Contributions

The main contributions of this work are summarized below:

1) We propose a unified, data-driven stochastic framework for generating realistic GPU utilization traces representative of modern AI datacenter workloads. The model integrates probabilistic formulations for capturing the diverse temporal and statistical features of AI compute clusters.
2) A complete datacenter power model is developed, linking IT load behavior to facility power consumption through thermal and Power Usage Effectiveness (PUE) relationships.
3) The generated traces are validated against empirical measurements from the MIT SuperCloud dataset.
4) All datasets, Python models, and simulation scripts are released as part of an open-source toolkit to facilitate benchmarking, system studies, and collaborative research on AI datacenter grid interactions.

II. Simulator Model

A modular approach is utilized to design the simulator. The datacenter loads include (i) IT Power dynamics (ii) Cooling power dynamics (iii) Ancillary services power. The server power data for datacenters can be obtained by utilizing open-source or proprietary data that capture real-world power consumption from heterogeneous computing clusters [9]. OALS is built such that if such data are available to the user, then it can be directly used in the simulator. However, such data are often not available and we create stochastic models to create data that emulate actual datacenter power consumption patterns. Fig.1 gives the overall structure of the simulator.

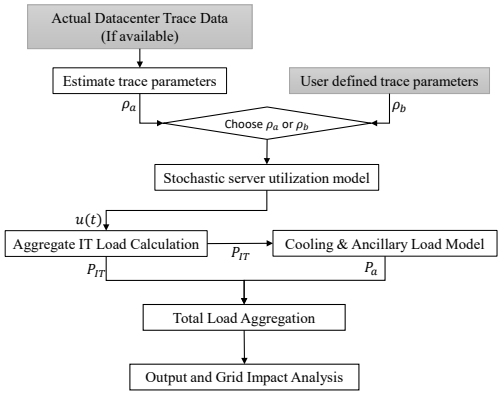

Fig. 1: OALS Simulator Overview

A. Load model for AI compute clusters

For the AI datacenters which handle the computation workloads, the IT load is dominated by GPU servers, with smaller contributions from CPUs, memory, and networking gear. These datacenters handle multiple complex AI tasks including training, inference, mini-batch processing and communication. The power consumption for each GPU is given by,

$$p_G(t) = p_{G,i} + (p_{G,m} - p_{G,i})u_G(t) \tag{1}$$

where, $p_{G,i}$ and $p_{G,m}$ are the idle and maximum power of the GPU and $u_G(t)$ is the GPU utilization at time t. The GPU utilization pattern is different for each of the different tasks handled by the datacenter. Table. II gives the characteristics associated with the different workloads handled by the server.

While these different workloads follow a certain pattern, the parameters defining these patterns are often stochastic. We propose stochastic models for each of the different types of workloads handled by GPUs. Thus, $u_G(t)$ in (1) can be defined using the models $u_1(t), u_2(t), u_3(t),$ and $u_4(t)$ depending on the type of workload as given below:

1) Steady inference/evaluation workloads: During deep learning inference tasks or batched inference process, the GPU remains active servicing requests (steady utilization) with occasional drops during queue-empty events, I/O stalls, or model reloads. This can be modeled as a shot-noise process with Poisson arrival times superimposed on a Gaussian mean level. That is,

$$u_1(t) = \mu_H - A_1 \sum_k \mathbb{1}_{[T_k,\, T_k+D_k)}(t), \tag{2}$$

TABLE II: AI Datacenter GPU workload characteristics

Trace ID	Workload Type	Signature/ Temporal Pattern	Mean Util. (%)	Dominant Phases
T1	Continuous Inference / Evaluation	Nearly flat, sustained plateau with short dips	80–85	Minimal; compute-bound
T2	Distributed training and Gradient Synchronization	Irregular tall spikes separated by long idle gaps	<30 (avg)	Queue / Network Latency
T3	Mini-Batch Training (Moderate Batch)	Rhythmic, high-frequency oscillations	45–55	Compute ↔ I/O Alternation
T4	Full-Batch or Compute-Bound Training	Strongly periodic, high-amplitude waveform	70–90	Compute-Limited

where, μ_H is the mean utilization, and $\mathbb{1}_{[T_k,\, T_k+D_k)}(t)$ is an indicator function which takes a value 1 if $t \in [T_k, T_k+D_k)$ and 0 otherwise. A_1 and D_k denote the amplitude and duration of random dips which are drawn from a uniform distribution.The dip arrival times $\{T_k\}$ follows a Poisson process with rate λ.

2) Distributed training and gradient synchronization: For communication-triggered GPU activity or in distributed setups, GPUs alternate between high computation phase (due to matrix operations) and idle phase (gradient synchronization via all-reduce). The utilization can be modeled as an on-off renewal process with random pulse arrivals:

$$u_2(t) = \sum_k U_k \, \mathbb{1}_{[T_k,\, T_k+D_k)}(t), \tag{3}$$

where inter-arrival times follow a poisson distribution, service durations D_k and U_k follow a general distribution (Eg. Uniform). This captures sparse, high-intensity bursts typical of asynchronous inference or gradient exchange.

3) Mini-batch training: Training workloads are quasi-periodic, where each cycle includes a high phase (with forward pass, backward pass matrix operations), and short synchronization (I/O or CPU-GPU coordination pauses between batches) steps. Each iteration (cycle) of length T_n is decomposed into an active fraction ϕ_n:

$$u_3(t) = \begin{cases} U_H + \eta_H(t), & t \in [C_n,\, C_n + \phi_n T_n), \\ U_L + \eta_L(t), & t \in [C_n + \phi_n T_n,\, C_{n+1}), \end{cases} \tag{4}$$

The uncertainty or variability within each phase is characterized by zero-mean perturbations $\eta_H(t) \in \mathcal{N}(0, \sigma_H^2)$ and $\eta_L(t) \in \mathcal{N}(0, \sigma_L^2)$. The cycle period and duty cycle are also taken as random values, $T_n \sim \mathcal{N}(\bar{T}, \sigma_T^2)$ and $\phi_n \sim \mathrm{Beta}(a, b)$.

4) Compute-Bound Training (Full-Batch Regime): Highly optimized training runs exhibit periodic high duty cycles corresponding to forward, backward, and optimizer steps. The dynamics should have a consistent amplitude and frequency to replicate the repetitive compute pattern rather than stochastic load. This pattern can be modeled using a low-jitter sinusoidal approximation given by:

$$u_4(t) = u_0 + A \sin(2\pi f t + \phi) + \varepsilon(t), \tag{5}$$

where, u_0 is the base utilization, A is the change drawn from a gaussian distribution, and frequency $f = 1/\bar{T}$ reflects the iteration rate producing the characteristic periodic utilization pattern of compute-saturated GPUs.

5) Aggregate Cluster Power: For a cluster with N GPUs running heterogeneous workloads, the total IT power is,

$$P_{\mathrm{IT}}(t) = \sum_{G=1}^{N_G} p_G(t), \tag{6}$$

B. Cooling and Ancillary Services Load Model

The cooling load is determined by the Power Usage Effectiveness (PUE) which is a ratio of total energy used by the facility and the energy used by the IT equipment, such as servers and storage. A lower PUE indicates greater efficiency, as it means a smaller portion of energy is being used for non-IT functions like cooling and power distribution. The non-IT or ancillary power of the datacenter is given by,

$$P_{\mathrm{a}}(t) = \beta(t)\, P_{\mathrm{IT}}(t), \tag{7}$$

where, β is the PUE of the datacenter. In the proposed model The PUE, β varies with IT loading to reflect cooling system efficiency given by,

$$\beta(t) = \beta_0 + k \frac{P_{IT}(t)}{P_{IT,0}} \tag{8}$$

where k is the sensitivity coefficient and $P_{IT,0}$ is the rated IT capacity. The final combined load model is given by:

$$P_{\mathrm{DC}}(t) = P_{IT}(t) + P_a(t) \tag{9}$$

The above open source model can be accessed from our github repository [10].

III. Simulator Use Cases

In this section, we discuss two use cases for the simulator. In the first case we compare multiple GPU traces with actual GPU traces obtained from the MIT supercloud dataset and in the second case study we use the simulator for grid impact analysis using a 39-bus New England test system.

A. Validation Case Study

To validate the traces produced by the proposed stochastic load models, we compare the synthesized GPU power traces against empirical data from the MIT SuperCloud dataset. This dataset is a publicly available repository of GPU and CPU telemetry collected from large-scale AI training and inference workloads. The MIT SuperCloud traces include measurements of GPU utilization, power, and memory bandwidth for thousands of training jobs executed on heterogeneous clusters, making them suitable as a reference benchmark.

Each synthetic trace type (T1–T4) is compared with corresponding workload segments in the MIT dataset exhibiting similar operational patterns (e.g., steady inference, bursty inference, mini-batch training, and communication-limited phases). The parameters used are given in Table.III

TABLE III: Simulation parameters for GPU utilization

Model	Parameter	Nominal Value / Range
T1	$\{\mu_H, \sigma, \lambda\}$	$\{0.85, 0.02, 0.02s^{-1}\}$
	D_k	3-10 s
	A_k	Uniform(0.1, 0.4)
T2	D_k	3-10 s
	U_k	Uniform(0.8, 1.0)
	λ	$0.02\,\mathrm{s}^{-1}$
T3	$\{\bar{T}, \sigma_T, U_H, U_L\}$	$\{10\mathrm{s}, 1\mathrm{s}, 0.7, 0.2\}$
	ϕ_n	Uniform(0.6, 0.8)
	η_H, η_L	$\mathcal{N}(0, 0.05^2)$
T4	$\{A, u_0, f\}$	$\{0.25, 0.75, 8s^{-1}\}$
	Noise	$\mathcal{N}(0, 0.03^2)$

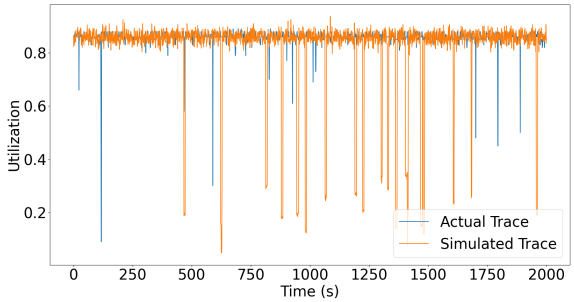

Fig. 2: Actual and simulated traces for T1

Fig.2 shows the actual and simulated traces for the continuous evaluation workload. It can be seen that the simulated trace effectively replicates the actual trace dynamics. Similar traces are obtained for all the different workloads.

The similarity between the simulated and actual traces are quantified using multiple statistical comparison metrics. Autocorrelation Function (ACF) similarity is used to capture the temporal persistence of fluctuations and it is given by the normalized ℓ_2 distance between ACFs, $e_{ACF} = \|\rho_{\mathrm{sim}}(\tau) - \rho_{\mathrm{MIT}}(\tau)\|_2$, captures temporal persistence of load fluctuations. Information loss between the actual and simulated traces is analyzed using the Kullback–Leibler (KL) divergence, $\mathrm{D_{KL}}$ and distributional similarity is analyzed using the two-sample Kolmogorov–Smirnov distance, $\mathrm{D_{KS}}$.

The values in Table.IV demonstrate that the proposed stochastic models closely replicate the statistical and temporal behavior of actual GPU traces. The low autocorrelation error values ($e_{\mathrm{ACF}} < 0.1$ for T1 and T2) indicate that the simulated traces successfully reproduce the short- and medium-range temporal dependencies present in steady and event-driven workloads. The corresponding KL-divergence values confirm that the probability distributions of utilization for these workloads are statistically consistent with the empirical data,

TABLE IV: Comparison between actual and simulated traces

Workload Type	e_{ACF}	$\mathrm{D_{KL}}$	$\mathrm{D_{KS}}$
T1	0.0878	3.0274	0.8795
T2	0.0765	2.4508	0.1466
T3	0.1644	0.5870	0.1790
T4	0.2011	1.7187	0.2320

while the KS-statistics ($D_{\mathrm{KS}} < 0.2$ for most cases) further validate their distributional similarity.

The preprocessing-limited workload (T4) exhibits comparatively higher errors across all metrics, primarily due to its irregular burst structure and asynchronous idle periods, which are more difficult to represent using low-order stochastic kernels. A visual inspection reveals that the dynamics are similar and thus they can be used as representative traces for power system studies.

B. Case Study 2: Grid Impact Analysis

To analyze the impacts of AI datacenters on the grid, we design a test system based on the 39-bus 3 area New England test system. The dynamic model parameters like generator, machine and governor parameter values are obtained from [11], [12]. The system is modified by placing two datacenters in area-1 at buses 26 and 27 as shown in Fig.3. The datacenter load is approximately 30% of the total area loads.

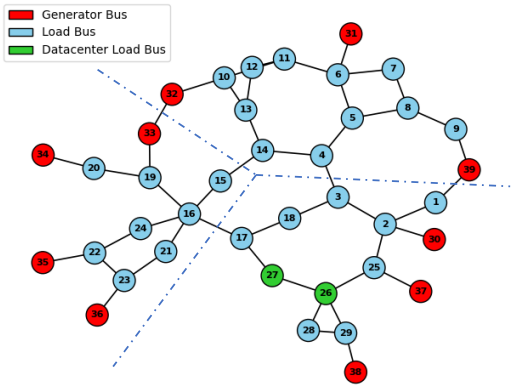

Fig. 3: Modified 39-bus New England system with DC loads

Each datacenter load is modeled using the composite stochastic framework described in Section II, comprising random combinations of the workload types (T1–T4). The resulting facility power trace, shown in Fig.4, exhibits both low-frequency and high-frequency variations arising from heterogeneous AI workloads. The frequency response is obtained using a detailed differential algebraic equations based dynamic model of the 3-area test system [13] simulated in MATLAB.

The system frequency response under this scenario is presented in Fig.5. Initially, the network operates under nominal loading conditions with tightly bounded frequency oscillations. At $t = 200$ s, the two datacenters are connected to the grid. It is

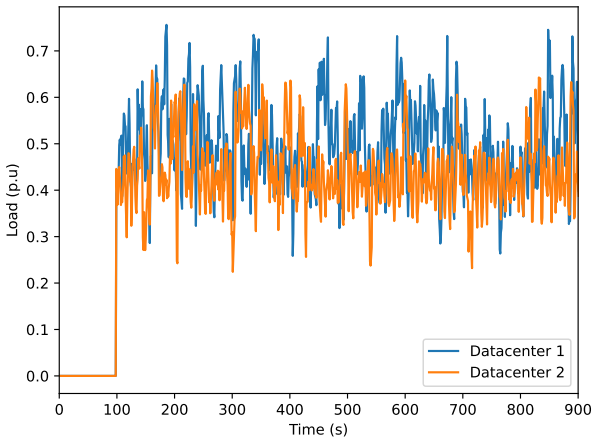

Fig. 4: Datacenter loads at 2 buses

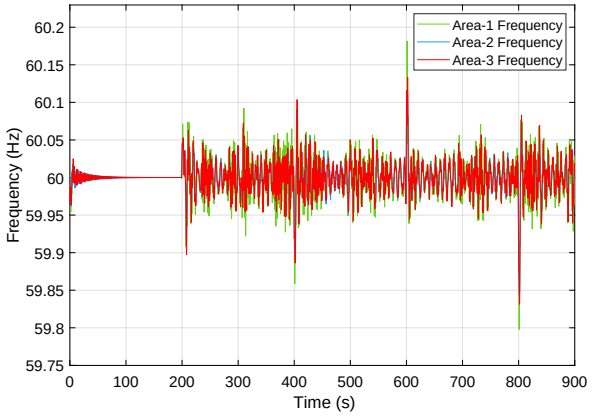

Fig. 5: System Frequency with datacenter loads

evident from the figure that on connecting the datacenter loads, the system experiences a noticeable increase in steady-state stress, characterized by larger amplitude and higher-frequency oscillations. The AI datacenter introduces oscillations in frequency with a nadir of 59.8Hz and the highest RoCoF of -0.1Hz. These values could worsen with under situations where the oscillations are additive. The oscillatory behavior arises due to the stochastic, burst-driven power variations of the AI workloads and the limited inertial contribution datacenter loads. This behavior highlights the potential of large-scale AI compute centers to modify the low-inertia dynamics, demanding advanced control and grid-support strategies for stable integration in future power systems.

The results demonstrate the value of the proposed OALS. It is important to note that the simulator is extendable to different data centers types by adapting the components and their stochastic behavior in the model described in Section II. Thus, OALS is a general framework that can be used by both the research community as well as utilities to advance the planning, operation and integration of various types of data centers into the grid.

IV. Conclusion

This paper presents an open-source simulation framework for modeling electrical behaviors of AI DCs and analyzing their impacts on power system dynamics. The proposed approach bridges high-performance computing workload characteristics with grid-interactive load modeling through a hierarchy of stochastic, thermal, and converter-level representations.

Validation against real-world GPU traces confirms the model's ability to reproduce the temporal signatures and statistical profiles of diverse AI workloads with high fidelity. When integrated into a benchmark power system, the simulations reveal that AI datacenters can amplify frequency oscillations and impose additional stress on conventional generation controls due to their high variability and low inertia.

The presented framework provides a scalable and transparent model for studying large-scale AI infrastructure in future low-inertia grids. This is the only open access generic AI data centers load simulator currently available to the power systems research community. The base components can be modified to develop more realistic and customizable models for system simulation and impact analyses. The alpha version of the simulator is currently available for public use and the team plans to continue developing follow on versions in the future. This includes incorporating adaptive converter controls, storage and UPS modeling and power-water aggregated models to extend the simulator's applicability to transmission-level planning and interdependent infrastructure resilience analysis.

References

[1] C.-L. He, "Real-time power management in virtualized data centers using predictive analytics," *International Journal of Cloud Computing and Database Management*, 2024.

[2] H. Wang, J. Huang, X. Lin, and H. Mohsenian-Rad, "Proactive demand response for data centers: A win-win solution," *Ieee Transactions on Smart Grid*, 2016.

[3] J. Li and W. Qi, "Toward optimal operation of internet data center microgrid," *Ieee Transactions on Smart Grid*, 2018.

[4] X. Hao, P. Liu, and Y. Deng, "Joint optimization of operational cost and carbon emission in multiple data center micro-grids," *Frontiers in Energy Research*, 2024.

[5] J. Sun, S. Wang, J. Wang, and L. Tolbert, "Development of a converter-based data center power emulator," *2021 IEEE Applied Power Electronics Conference and Exposition (APEC)*, pp. 126–133, 2021.

[6] ——, "Dynamic model and converter-based emulator of a data center power distribution system," *IEEE Transactions on Power Electronics*, vol. 37, pp. 8420–8432, 2022.

[7] D. A. Kez, A. Foley, S. Muyeen, and D. Morrow, "Manipulation of static and dynamic data center power responses to support grid operations," pp. 182 078–182 091, 2020.

[8] Y. Zhang, H. Tang, H. Li, and S. Wang, "Integration and interaction of next-generation ai-focused data centers with smart grids and district energy systems: The state-of-the-art, opportunities and challenges," 2025.

[9] S. Samsi, M. L. Weiss, D. Bestor, B. Li, M. Jones, A. Reuther, D. Edelman, W. Arcand, C. Byun, J. Holodnack, M. Hubbell, J. Kepner, A. Klein, J. McDonald, A. Michaleas, P. Michaleas, L. Milechin, J. Mullen, C. Yee, B. Price, A. Prout, A. Rosa, A. Vanterpool, L. McEvoy, A. Cheng, D. Tiwari, and V. Gadepally, "The mit supercloud dataset," in *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, 2021, pp. 1–8.

[10] S. Amulya, "Oals," https://github.com/power-grid-cyber, 2025.

[11] H. Bevrani, *Robust power system frequency control*. Springer, 2009.

[12] A. Amulya, K. S. Swarup, and R. Ramanathan, "Multivariate spectral analysis and hypothesis testing-based robust attack detection for multiarea frequency control," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–13, 2023.

[13] K. Prabha, *Power System Stability and Control*. McGraw-Hill, 1995.