

Grid Frequency Stability Support Potential of Data Center: A Quantitative Assessment of Flexibility

Pengyu Ren*, Wei Sun*, Yifan Wang*, Gareth Harrison*

*School of Engineering, University of Edinburgh, Edinburgh, Scotland

Email: s2192464@ed.ac.uk, w.sun@ed.ac.uk, s2154060@ed.ac.uk, gareth.harrison@ed.ac.uk

Abstract—The rapid expansion of data center infrastructure is reshaping power system dynamics by significantly increasing electricity demand while also offering potential for fast and controllable flexibility. To ensure reliable operation under such conditions, the frequency-secured unit commitment (Safe UC) problem must be solved with enhanced modeling of demand-side frequency response. In this work, we propose a data-driven linearization framework based on decision tree-based constraint learning (DT-CL) to embed nonlinear nadir frequency constraints into mixed-integer linear programming (MILP). This approach enables tractable co-optimization of generation schedules and fast frequency response (FFR) from data centers. Through case studies on both a benchmark system and a 2030 future scenario with higher DC penetration, we demonstrate that increasing the proportion of flexible DC load consistently improves system cost efficiency and supports renewable integration. However, this benefit exhibits diminishing marginal returns, motivating the introduction of the Marginal Flexibility Value (MFV) metric to quantify the economic value of additional flexibility. The results highlight that as DCs become a larger share of system load, their active participation in frequency response will be increasingly indispensable for maintaining both economic and secure system operations.

Index Terms—data center, power system, frequency response, unit commitment.

I. INTRODUCTION

The rise of artificial intelligence (AI), big data, and cloud computing has significantly driven the growth of data centers in recent years. According to the International Energy Agency (IEA), global data center electricity consumption reached 415 TWh in 2024—already 132% of the UK’s total electricity use that year—and is projected to more than double to 945 TWh by 2030 [1]. This explosive growth has raised global concerns over the sustainability and grid compatibility of data center operations, especially under net-zero carbon targets. Unlike conventional residential loads, the rapid pace and geographic concentration of data center development poses unique challenges for local electricity infrastructure [2].

Amid this trend, increasing attention has been given to the flexibility potential of data centers—the ability to modulate workload and electricity demand in response to grid signals. [3] shows that AI-focused, GPU-heavy high-performance computing (HPC) data centers can provide power system flexibility at significantly lower cost than traditional CPU-based HPC centers, highlighting their potential as cost-effective resources for grid balancing. Figure 1 illustrates recent developments in data center flexibility. Current strategies include dynamically shifting workloads in time and location, coordinating resource

allocation across multiple cloud service providers, and aligning computing tasks with renewable energy availability.

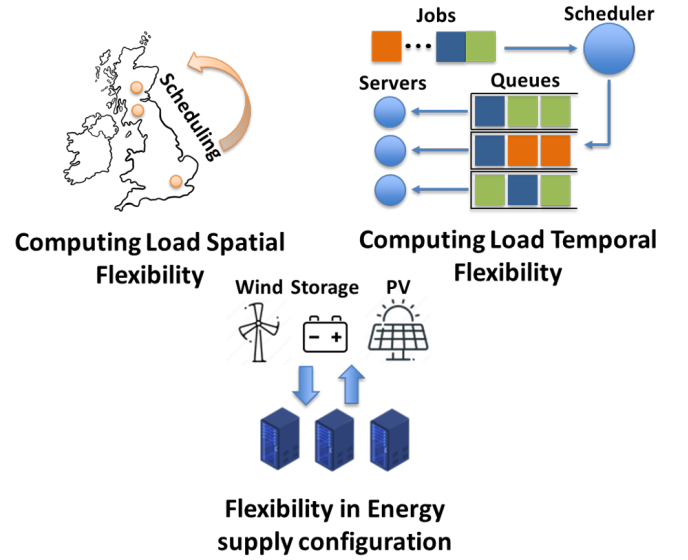


Fig. 1: Strategies for enabling data center flexibility

In recent years, a growing body of literature has explored both the temporal and spatial flexibility of data centers, recognising their potential to provide grid support services such as load shifting [4] and fast frequency response [5]. For example, [6] investigates how data centers can reduce their total electrical consumption costs through workload scheduling while maintaining Quality of Service (QoS) guarantees. Similarly, [7] proposes an energy-efficient job scheduling algorithm in cloud data centers by classifying jobs and applying preemption policies to maximize active host utilization and minimize the number of active physical machines, achieving up to 46% energy savings compared to non-energy-aware baselines. [8] proposes a privacy-preserving federated reinforcement learning framework for collaborative job scheduling across cloud providers, enabling energy sharing through decentralized decision-making while protecting operational privacy. [9] proposes a spatio-temporal workload migration mechanism that reduces carbon emissions by flexibly shifting tasks across geographically distributed data centers to match multi-regional renewable energy availability.

However, most of these studies adopt the perspective of

data center operators, focusing on minimizing electricity bills or improving internal energy efficiency through workload scheduling [6]–[8]. A few recent works have begun to explore the system-level benefits of spatial and temporal workload shifting, such as improving renewable energy utilization [9], [10], reducing grid violations [11], and lowering carbon emissions [12]. Nevertheless, these efforts typically rely on simulation-based analyses and do not integrate such flexibility into formal grid operation models. In particular, there remains a lack of security-aware dispatch models that assess the impact of data center flexibility on frequency response and unit commitment decisions in real power systems.

To better understand the system-wide value of data center flexibility, it is crucial to shift the analytical perspective from isolated, facility-level optimizations to a holistic view of integrated power system operations [13]. A central challenge in this context lies in coordinating flexible loads—such as data centers—within established power system scheduling frameworks, particularly the Unit Commitment (UC) problem. In UC, the system operator determines the optimal dispatch and on/off status of generation units to minimize total operating costs. When frequency constraints are considered, the model must also ensure sufficient system inertia, typically requiring the commitment of certain thermal generators [14].

Several studies have extended the UC framework to incorporate frequency-related considerations. For instance, [15] proposes a stochastic UC model that integrates frequency constraints to assess the economic impact of frequency response capabilities. Similarly, [16] introduces a stochastic planning model that incorporates the frequency response behavior of wind power to ensure system frequency stability. In parallel, [17] explores the role of large-scale data centers in the 2030 Irish power system, embedding their operations into a mixed-integer unit commitment formulation to study cost and flexibility trade-offs.

However, many existing models oversimplify frequency response dynamics, especially when dealing with multi-resource frequency response like data center. Homogeneous response characteristics across all resources are often assumed, overlooking critical device-specific factors such as activation delays, ramping behavior, and capacity constraints. These limitations may lead to inaccurate assessments of flexible resources’ true value and hinder the development of effective frequency support strategies. To fill in the above research gap, this paper proposes an innovative safe UC method based on decision tree linearization. In detail, the main contributions of this paper are summarized as follows:

- Quantitatively evaluate the value of data center frequency flexibility in improving the reliability of the power system under high penetration of renewables energy. By enabling data centers to provide fast frequency response (FFR), the proposed framework reduces overall system operation costs and mitigates reliance on conventional thermal generators, thereby supporting a more efficient and sustainable power system operation.

- Propose a safety-aware unit commitment (Safe UC) model that integrates data-driven frequency safety constraints derived from historical operation data. We introduce a novel constraint generation mechanism using a logistic-regression-based slope tree, which partitions the operational space into safe and unsafe regions. The resulting piecewise linear inequalities are directly embedded in the MILP formulation, enabling interpretable and computationally tractable enforcement of empirical safety boundaries.
- Demonstrate the effectiveness of the proposed approach on a high-renewable test system. Results show that incorporating data center frequency response improves renewable utilization, reduces curtailment, and contributes to secure operation under increasing system uncertainty, offering practical insights for future grid planning and data center participation.

II. SAFE UNIT COMMITMENT MODELLING

To address the challenges of integrating flexible data center demand into the UC problem, we propose a hybrid framework that combines traditional optimization modeling with data-driven decision-making. As illustrated in Fig. 2 our method augments a conventional UC formulation with a decision tree (DT)-based constraint learning mechanism to improve operational safety and adaptability under uncertainty.

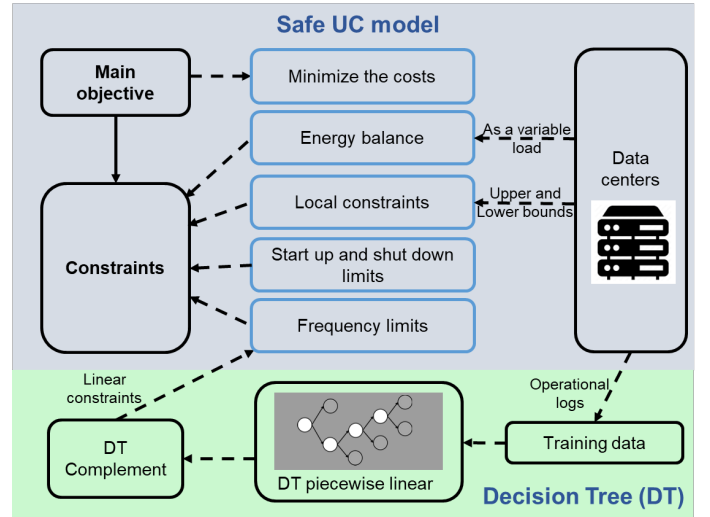


Fig. 2: Outline of Safe UC modelling.

The Safe UC model lies at the core of the framework. Its primary objective is to minimize total generation cost, subject to a range of physical and operational constraints including energy balance, frequency limits, local network limitations, and start-up/shut-down dynamics. Data centers are modeled as flexible and controllable loads, whose consumption is treated as a variable within specified upper and lower bounds. These bounds, derived from operational capabilities, are incorporated into the local constraints of the UC formulation.

To enhance safety and robustness, we introduce data-driven safety constraints learned from historical operational logs. Specifically, we train a decision tree classifier on past system data to identify safe vs. unsafe operating regions. The resulting tree is then converted into a piecewise linear form, which is further translated into a set of linear constraints compatible with mixed-integer linear programming (MILP) formulations.

The DT complement module, as shown in the lower part of Fig. 2, serves as a constraint-generation engine. It translates empirical knowledge into constraints that can be seamlessly combined with the analytical UC model. In doing so, the overall framework maintains interpretability while leveraging real-world data to improve decision quality, particularly under scenarios involving high demand variability and renewable generation uncertainty.

A. Defining data center flexibility

Before formulating the safe unit commitment model, it is essential to define the nature of flexibility that data centers can provide to the power system. As discussed in the introduction part, numerous studies have demonstrated that data centers possess inherent flexibility due to their internal structure and the nature of computing workloads they handle.

The degree of controllability largely depends on the types of incoming jobs, which can be broadly divided into:

- **Delay-sensitive (real-time) workloads**, which must be executed immediately upon arrival to meet user expectations or quality-of-service requirements.
- **Delay-tolerant (batch) workloads**, which can be scheduled flexibly within a given time window, provided they complete before a specified deadline.

In general, batch workloads are the primary source of demand-side flexibility in data centers. These tasks can be modulated, deferred, or even interrupted without violating service-level agreements. This flexibility enables data centers to adapt their power consumption in response to system needs.

In this study, since our focus is on enabling FFR, only workloads that can be rapidly deferred—typically batch jobs with high schedulability—are considered. Other forms of flexibility, such as long-term workload peak shifting or modulation of cooling system loads, are beyond the scope of this work and not included in the model.

B. Objective function

The objective of the unit commitment model with fast frequency response is to minimise the total operating cost over a scheduling horizon:

$$\min \sum_{t \in \mathcal{T}} \left[\underbrace{\sum_{i \in \mathcal{I}} (\beta_1 P_{i,t}^2 + \beta_2 P_{i,t} + \beta_3)}_{\text{Generator costs}} + \underbrace{\gamma_1 R_t^{\text{DC}} + \gamma_2 (R_t^{\text{DC}})^2}_{\text{DC FFR costs}} \right] \quad (1)$$

where $P_{i,t}$ is the active power output of generator i at time step t , R_t^{DC} is the DC fast frequency response potential at

time t , $\beta_1, \beta_2, \beta_3$ are generation cost coefficients, γ_1, γ_2 are linear and quadratic cost coefficients for DC response.

The formulation balances two key components of system operation: the economic efficiency of power generation and the provision of frequency regulation services by flexible demand resources. The generator cost term reflects the traditional economic dispatch, where generation units are scheduled based on their cost curves. In contrast, the DC FFR cost term quantifies the economic burden of engaging data centers in frequency support, penalizing excessive reliance on demand flexibility through both fixed and escalating marginal costs.

Importantly, the inclusion of R_t^{DC} as a decision variable bridges the physical UC optimization with the data-driven safety constraints. These variables are not only subject to operational bounds, but are also constrained by learned safe-operating regions derived from historical data via the DT mechanism.

This coupling ensures that the optimisation respects both physical grid limits and empirically derived safety margins, enhancing the realism and robustness of the scheduling decisions. As such, the objective function lies at the core of a hybrid paradigm that marries analytical modeling with machine learning-enabled constraint generation, enabling secure and cost-effective operation under uncertainty from renewable generation and flexible loads.

C. Conventional Unit commitment constraints

In addition to the frequency considerations discussed later, the core unit commitment (UC) model is governed by a set of traditional operational constraints that ensure the physical feasibility and reliability of the system. These include power balance, generator capacity limits, ramping limitations, and minimum up/down time requirements, as described below:

$$\sum_i P_{i,t} + \sum_i P_{it}^{\text{wind}} = P_t^{\text{curt}} + P_t^{\text{DC}} + P_t^{\text{load}} \quad (2)$$

$$P_i^{\min} u_{i,t} \leq P_{i,t} \leq P_i^{\max} u_{i,t} \quad (3)$$

$$P_{i,t} - P_{i,t-1} \leq \rho_i^{\text{up}} u_{i,t} \quad (4)$$

$$P_{i,t-1} - P_{i,t} \leq \rho_i^{\text{up}} u_{i,t} \quad (5)$$

$$1 - u_{i,t} \geq \sum_{d=1}^{\delta_i} (1 - u_{i,t-d}) \quad (6)$$

where P_{it}^{wind} represents wind power generation; P_t^{curt} is the curtailed power; P_t^{DC} is the power consumed by data center loads; P_t^{load} is the total system demand; P_i^{\min} and P_i^{\max} are the minimum and maximum generation limits of unit i , respectively; ρ_i^{up} and ρ_i^{up} denote the generator's ramp-up and ramp-down limits; and δ_i is the minimum down-time duration required once unit i is turned off.

Each constraint serves a specific operational or stability requirement. Equation (2) ensures active power balance between generation and demand (including curtailed and DC load). Equation (3) limits generator outputs based on commitment status. Equation (4)-(5) enforce generator ramp-up and ramp-down rate limits. Equation (6) ensures that once a generator

is turned off, it must remain off for a minimum down-time duration δ_i .

D. Frequency response modelling

While the above constraints define the traditional operational boundaries of the UC problem, they do not explicitly capture the system's dynamic behavior during frequency disturbances. In modern grids with high renewable penetration and demand-side flexibility, modelling frequency response characteristics becomes essential to ensure stability and system security [18]. In this context, we incorporate an explicit model of frequency response behavior from both conventional generators and data center loads, as described in the next section.

Suppose a power outage or unit dropout occurs, when the system starts to shift in frequency due to a power imbalance. When the frequency offset reaches the frequency deadband Δf_{DB} set by the system, the ancillary service starts to respond. As shown in Fig. 3, The model assumes that the fast frequency response process of both data center and conventional generators is a ramp function, and they do not start responding at the same time.

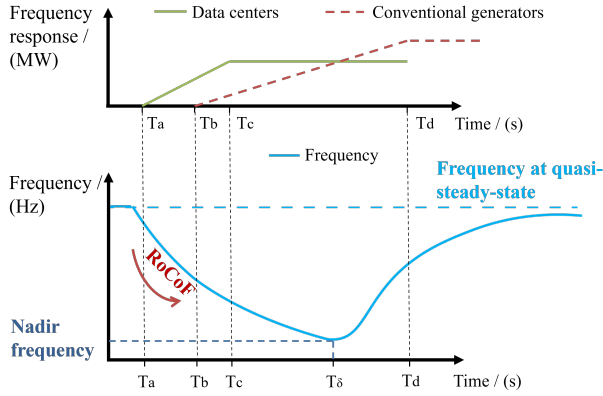


Fig. 3: Assumed frequency response process of data center and conventional generators.

The frequency constraint of the system can be delineated into three distinct components, as the Fig. 3 shows. The initial component imposes a constraint on the Rate of Change of Frequency (RoCoF) during a frequency variation, stipulating that the rate of change must not be excessively rapid to avert excessive stress on transmission lines. The second component encompasses a constraint on the frequency at nadir, ensuring that the system's frequency remains stable within an acceptable range. The final component comprises a constraint on the frequency at which the system attains a quasi-steady state following its frequency response.

Commencing at time T_a , the frequency offset attains the deadband frequency threshold, prompting the DCs to initiate a ramping-type FFR. At time T_b , the conventional generator commences its FFR performance. At T_c , the FFR of the DC unit achieves its stable capacity and enters a maintenance phase. Subsequently, at T_d , the FFR of the conventional

generator also reaches its stable capacity. This process is shown in Fig. 3. The ramping functions for the two are shown below:

$$\Delta R_{DC}(t) = \begin{cases} 0 & \text{if } t < T_a \\ \frac{R_{DC}^{sta}}{T_c - T_a} \times (t - T_a) & \text{if } T_a \leq t \leq T_c \\ R_{DC}^{sta} & \text{if } t \geq T_c \end{cases} \quad (7)$$

$$\Delta R_{cg}(t) = \begin{cases} 0 & \text{if } t < T_b \\ \frac{R_{cg}^{sta}}{T_d - T_b} \times (t - T_b) & \text{if } T_b \leq t \leq T_d \\ R_{cg}^{sta} & \text{if } t \geq T_d \end{cases} \quad (8)$$

where the $\Delta R_{DC}(t)$ and the $\Delta R_{cg}(t)$ are FFRs of the DCs and conventional generators (basically are thermal units). The R_{DC}^{sta} and the R_{cg}^{sta} are maximum FFRs could be provided. T_a and T_b are the start times of the FFRs of the DCs and conventional generators, respectively. Meanwhile, T_c and T_d are the times at which the DCs and conventional generators reach their maximum FFRs, respectively.

E. Rate of change of frequency limit

In addition to modelling the ramping behavior of flexible resources, a critical aspect of frequency security lies in constraining the Rate of Change of Frequency (RoCoF) immediately following a contingency. RoCoF is a dynamic metric that reflects how rapidly the system frequency changes in response to power imbalances and is inversely proportional to the system's total inertia. Excessive RoCoF may lead to protection system misoperations, equipment damage, or system instability.

The system inertia H_s can be expressed as:

$$H_s = \frac{\sum_{g \in \mathcal{G}} H_g \times P_g^{\max} \times B - \Delta P_L^{\max} \times H_L^{\max}}{f_0} \quad (9)$$

where H_s is the system inertia, H_g is the inertia constant of each generator, P_g^{\max} is the maximum output of each generator, B is the binary control variable, ΔP_L^{\max} is the maximum loss of load, H_L^{\max} is the inertia constant of load, and f_0 is the nominal frequency.

Accordingly, the system-wide RoCoF is formulated as:

$$\delta = \frac{\Delta P_L^{\max}}{2H_s} \leq \delta_{\max} \quad (10)$$

where δ denotes the instantaneous rate of frequency change and δ_{\max} is the permissible upper limit to ensure safe system operation.

F. Nadir Frequency Constraint

While the RoCoF constraint governs the initial rate of frequency deviation, frequency security must also account for the lowest frequency point reached after a disturbance—commonly referred to as the nadir. Following a loss of generation, the frequency continues to drop even after primary response is triggered, until sufficient energy is injected to restore balance. The nadir frequency thus depends on three

key factors: system inertia, total primary frequency response, and the magnitude of the power imbalance.

To avoid triggering under-frequency load shedding (UFLS) or protection trips, a nadir constraint is imposed to ensure that system frequency stays above a critical threshold during transients. The time evolution of system frequency $\Delta f(t)$ can be described using the linearised swing equation [19] :

$$2H_s \frac{d\Delta f(t)}{dt} + D \cdot P_D \cdot \Delta f(t) = \sum_{g,s \in \mathcal{G}, \mathcal{S}} \Delta P_{g,s}(t) - \Delta P_L^{\max} \quad (11)$$

where D is the frequency damping coefficient, P_D is the total demand, and $\Delta P_{g,s}(t)$ is the frequency response power from resource g of type s .

By substituting the ramping expressions of $\Delta R_{DC}(t)$ and $\Delta R_{cg}(t)$ into this differential equation and solving for $\Delta f(t)$, we derive a four-phase nonlinear frequency trajectory. The nadir point, influenced by the interplay of inertia and response ramping delays, does not admit a closed-form linear constraint in standard MILP formulations. Therefore, we reformulate the nadir constraint using a data-driven linear approximation, which is detail discussed in Section. III.

G. Quasi-Steady-State (QSS) Constraint

After the frequency nadir, the system enters a quasi-steady-state period where the frequency deviation is relatively constant. To maintain operational integrity and avoid secondary frequency control actions, the residual frequency deviation must not exceed a permissible value:

$$L_{\max} - R_t \leq \xi \cdot D^{\text{peak}} \cdot \lambda \quad (12)$$

where L_{\max} is the maximum disturbance size, R_t is the total available frequency response at time t , D^{peak} is peak demand, and ξ, λ are empirical QSS coefficients derived from operational data.

III. REFORMULATION OF NADIR FREQUENCY LIMIT

Considering that the frequency response characteristics of data centers differ from those of conventional generators, the original nadir frequency constraint becomes even more nonlinear. As a result, the nadir frequency limit cannot be directly incorporated into the MILP formulation of the MILP problem. To overcome this, we apply a data-driven linearization technique based on decision trees.

A. Decision tree linearization

A decision tree is a supervised learning algorithm that recursively partitions the input space into subsets based on feature values, ultimately forming a tree-like structure where each leaf node represents a prediction, as shown in Fig. 4a. When used for piecewise linearization of a nonlinear function, the decision tree algorithm divides the function's domain into multiple regions, as shown in Fig. 4b, each of which can be approximated by a linear model.

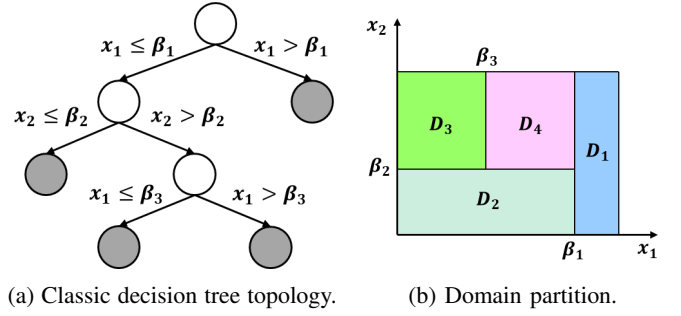


Fig. 4: An illustration on a simple decision tree with splitting knots.

DTs have been used in the power systems domain, including [20] who used a decision tree method to linearize the Q-V curve for voltage stability in power systems. In contrast, the present work conducts linearization in a three-dimensional space, due to the three-variable nature of the nonlinear frequency nadir constraints, which makes the process significantly more challenging.

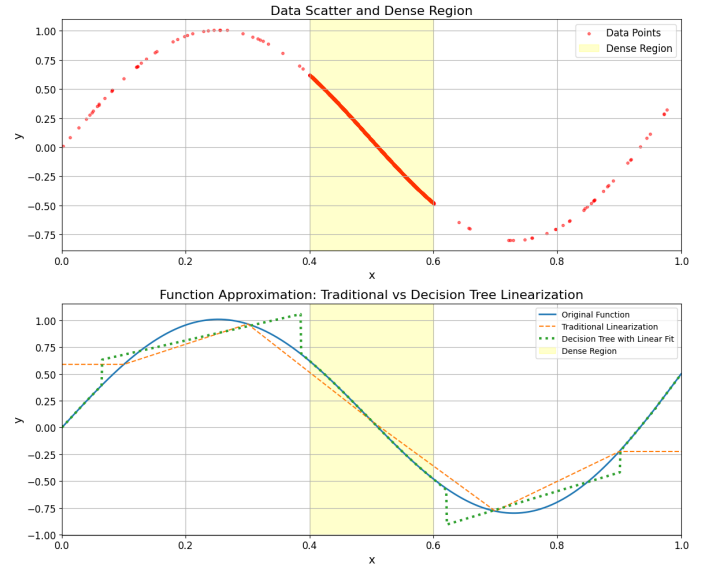


Fig. 5: Comparison between conventional piecewise linear and decision tree linear models.

Figures 5 illustrate the differences between the linearization method using a Decision Tree and the traditional piecewise linearization approach. The upper plot shows a non-uniform sample of a sine curve, where the yellow region represents the densely sampled data points. As shown in the lower plot, the conventional piecewise linearization method adopts equally spaced segments to minimize the overall deviation between the final linearized expression and the original nonlinear relationship. In contrast, the Decision Tree captures the characteristics of the data. With the same number of segments, it sacrifices accuracy in less dense regions to ensure lower errors in areas with high-frequency data.

This approach provides an adaptive method for piecewise linearization, as the decision tree automatically determines the optimal partitions based on the data. However, the accuracy of the approximation depends on the tree's depth and splitting criteria, which control the granularity of the partitioning. Overly shallow trees may lead to coarse approximations, while overly deep trees risk overfitting and unnecessary complexity. The training of the decision tree, including parameter selection and interval adjustment, will be explained in detail in the next subsection.

B. DT training algorithm

To enable the training of the decision tree classifier, we first construct a Simulink-based frequency response simulation framework based on the analytical nadir frequency equation derived earlier. This control system mimics the dynamic behavior of frequency deviation following generation loss and includes both generator inertia and flexible load responses from data centers. For each unit commitment solution, we extract relevant features including the power output of conventional generators, the scheduled load level of data centers, and other system operating conditions. These are used as input variables to the Simulink model, while different data center response times are configured as system parameters.

The output of each simulation instance is a binary safety label, indicating whether the corresponding operating point satisfies the nadir frequency constraint under a specific disturbance. Collectively, this process generates a labeled dataset suitable for supervised learning, where each input vector represents an operational scenario and the label denotes its safety status.

To encode safety constraints directly from historical operation data, we construct a decision tree whose internal nodes are defined by logistic regression classifiers, referred to as a slope tree. Unlike conventional regression or classification trees that use axis-aligned splits, this structure allows each node to introduce a linear decision boundary, yielding a more compact and flexible partitioning of the input space.

Given a labeled dataset (\mathbf{x}_i, y_i) , where \mathbf{x}_i encodes the system's operating features and $y_i \in \{0, 1\}$ indicates whether the operating point is safe or unsafe, the tree construction proceeds recursively as shown in Algorithm 1. At each node, we fit a logistic regression model

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + c$$

to separate the two classes. The dataset is then split based on the sign of the score $f(\mathbf{x})$: the left child receives points with $f(\mathbf{x}) < 0$, and the right child receives those with $f(\mathbf{x}) \geq 0$. The process stops when the maximum tree depth or minimum sample threshold is reached, or when all samples in the node belong to the same class.

The DT complement module translates empirical safety knowledge into linear inequality constraints, which can be seamlessly embedded into MILP formulations. Unlike conventional regression or classification trees that use axis-aligned splits, the slope tree is a form of oblique decision tree, in

Algorithm 1 BuildSlopeTree: Logistic Regression-Based Slope Decision Tree

Require: Feature matrix X , label vector y , current depth d , max depth d_{max} , minimum samples n_{min}

Ensure: Tree node or **None**

```

1: if  $d \geq d_{max}$  or  $|X| < n_{min}$  or all  $y_i$  are equal then
2:   return None
3: end if
4: Train logistic regression model  $f(x) = \mathbf{a}^\top \mathbf{x} + c$  on  $(X, y)$ 
5: Compute decision scores  $\mathbf{s} \leftarrow f(X)$ 
6: Partition data:
   • Left subset:  $X_L, y_L \leftarrow \{(\mathbf{x}_i, y_i) \mid s_i < 0\}$ 
   • Right subset:  $X_R, y_R \leftarrow \{(\mathbf{x}_i, y_i) \mid s_i \geq 0\}$ 
7: if  $X_L$  or  $X_R$  is empty then
8:   return None
9: end if
10: Recursively build left subtree:  $T_L \leftarrow$ 
    BuildSlopeTree( $X_L, y_L, d + 1, d_{max}, n_{min}$ )
11: Recursively build right subtree:  $T_R \leftarrow$ 
    BuildSlopeTree( $X_R, y_R, d + 1, d_{max}, n_{min}$ )
12: if  $T_L = T_R = \mathbf{None}$  then
13:   return None
14: end if
15: return Node  $(\mathbf{a}, c, T_L, T_R)$ 

```

Algorithm 2 CollectConditions: Traverse a slope tree to extract MILP-compatible linear inequalities

Require: Tree root node T

Ensure: List of linear inequalities (\mathbf{a}, c)

```

1: if  $T = \mathbf{None}$  then
2:   return empty list
3: end if
4: Initialize list  $L \leftarrow [(\mathbf{a}, c)]$ 
5: Append  $CollectConditions(T.left)$  to  $L$ 
6: Append  $CollectConditions(T.right)$  to  $L$ 
7: return  $L$ 

```

which each internal node applies a logistic regression classifier to introduce a linear (oblique) decision boundary. This enables more compact and flexible partitioning of the input space, better aligning with the geometry of the nonlinear nadir frequency constraint.

The adaptivity of the tree structure enables data-driven refinement of safety boundaries. However, the depth and splitting criteria must be carefully chosen to balance approximation accuracy and computational tractability. Further details on the training procedure and constraint encoding will be provided in the following section.

However, the output of the decision tree is not immediately suitable for optimization solvers. To incorporate its results into the unit commitment problem, the learned decision rules must be reformulated into MILP-compatible linear inequalities. This is accomplished by traversing the tree using the procedure in Algorithm 2, which collects the conjunction of node-level

conditions along each path to a safe leaf. The resulting set of inequalities defines the feasible operating region in a form that can be directly embedded into the MILP formulation.

C. Constraint embedding

The resulting convex relationship between system frequency nadir and control variables can be expressed as a linear inequality. Specifically, this data-driven convex surface can be approximated using supervised learning techniques, allowing it to be formulated as part of the MILP model:

$$\theta_1 R_t^{\text{gen}} + \theta_2 R_t^{\text{DC}} + \theta_3 H_t + \theta_4 \geq 0 \quad (13)$$

Here, θ_1 , θ_2 , θ_3 , and θ_4 are slope coefficients extracted via supervised learning using logistic regression-based decision trees. This transformation embeds the frequency nadir constraint as a linear inequality, enabling compatibility with MILP solvers used in unit commitment formulations.

By doing so, the nonlinear frequency security boundary is effectively captured by a set of interpretable and computationally tractable constraints, integrating physics-informed machine learning into power system scheduling.

IV. CASE STUDY AND ANALYSIS OF RESULTS

The power system dataset used in this case study is based on a modified IEEE 118-bus system [21], shown in Fig. 6, in which a high level of wind power penetration has been incorporated to simulate future power systems with renewable energy and data center integration. Several conventional generators have been replaced or augmented with wind farms located at strategically selected buses to reflect spatial diversity and variability. The optimization and solving language employed in this study is Julia 1.10.2.

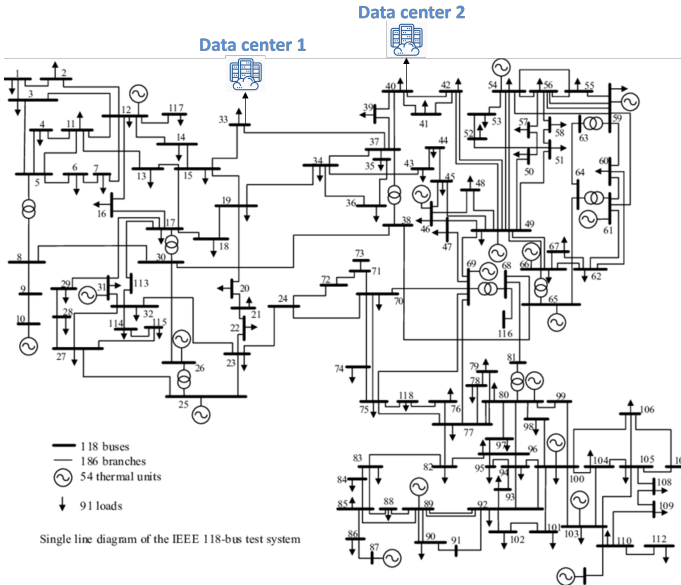


Fig. 6: Modified IEEE 118 bus test system [21], including two data centers with 500MW peak demand each.

The system under study is configured with a total demand of 5242 MW, including 1000 MW from data center loads. Table I summarizes the installed capacities of conventional and renewable generation units, as well as the demand components.

TABLE I: System Configuration Summary

Component	Capacity / Demand (MW)
Conventional Generators	5437.69
Wind Turbines	2718.84
Data Center Load	1000.00
Other System Load	4242.00
Total Load	5242.00

The 1000 MW load of data center represents the peak rated capacity of the data center component. In the simulation model, this value serves as the upper bound of flexible load. Rather than assuming constant demand, the actual data center power consumption at each hour is derived from a year-long operational log covering 8760 hours from the year 2019. The normalized load profile (α_t) is obtained from publicly available standard demand profiles developed by UK Power Networks, which provide hourly data for various demand types including data centers [22]. Specifically, the real-time load is modeled as a time-varying share of the peak, computed as:

$$P_{\text{DC},t} = \alpha_t \cdot P_{\text{DC}}^{\text{max}}, \quad \alpha_t \in [0, 1]$$

where $P_{\text{DC}}^{\text{max}} = 1000 \text{ MW}$ in the benchmark scenario, and α_t represents the normalized load profile obtained from historical records. This approach captures the temporal variability of computing workloads and ensures more realistic representation of data center operation in the system-level analysis.

A. Benchmark system cost result

We evaluate the impact of data center-based frequency response on system performance through sensitivity analyses that vary two key parameters: the proportion of flexible capacity and the speed of responses.

1) Proportion of Flexible Capacity within Data Centers:

The share of total data center load capable of participating in frequency response is varied from 0% (no participation) to 100% (full controllability). Greater flexibility generally reduces system operating costs by lowering the need for expensive spinning reserves or thermal ramping during frequency events. However, the marginal benefit declines once flexibility reaches a level sufficient to meet most frequency reserve requirements.

2) Response Speed of Data Center FFR:

We also assess the sensitivity of system cost to FFR delivery speed. Faster responses allow frequency deviations to be contained earlier, reducing reliance on slower and costlier primary frequency control from generators. Conversely, delayed responses diminish the effectiveness of data centers as substitutes for fast reserves.

As shown in Fig. 7, increasing DC flexibility consistently lowers system costs for all tested response times, with the largest savings achieved at lower flexibility levels (0–50%).

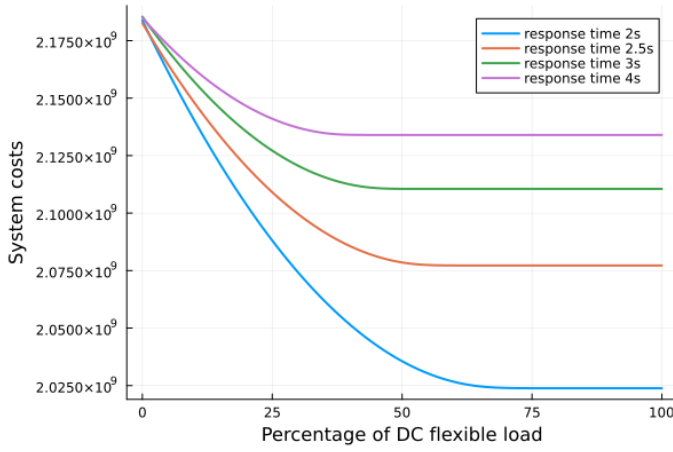


Fig. 7: Impact of the percentage of DC flexible load on system costs under different FFR response times.

Beyond 75%, the cost reduction plateaus, reflecting saturation of available demand-side flexibility. Faster response times deliver noticeably greater savings than slower ones, underscoring the importance of response latency in unlocking the full economic value of DC-based FFR.

To quantify the marginal economic benefit of increased data center flexibility, we define the *Marginal Flexibility Value* (MFV) as the system cost reduction per 1% increase in DC flexible load capacity:

$$\text{MFV} = \frac{C_i - C_j}{\phi_j - \phi_i}$$

where C_i and C_j are the total system costs corresponding to data center flexible load proportions ϕ_i and ϕ_j , respectively. The variable $\phi \in [0, 1]$ denotes the share of the data center load that is responsive to grid frequency signals, expressed as a fraction of the total data center capacity. The unit of MFV is million dollars per 1% flexible capacity (\$M/%), representing the cost reduction achieved by enabling one additional percent of data center load to provide frequency response.

Fig. 8 presents the MFV values under different FFR response times. The initial segments (0–25%) exhibit the highest MFV, particularly at faster response times, indicating that early investments in fast-response flexibility provide the greatest economic return. As flexible load proportion increases, MFV declines, reflecting diminishing marginal savings due to saturation of the system’s frequency reserve needs.

Overall, the results indicate that both the depth (i.e., the proportion of controllable load) and the speed (i.e., response time) of DC-based FFR are key drivers of system cost savings. System operators and policymakers may therefore consider incentivizing not only participation levels but also technological upgrades that shorten response times.

B. 2030 Scenario with Increased Data Center Load Penetration

To evaluate the future implications of growing data center demand, we construct a 2030 scenario based on projected

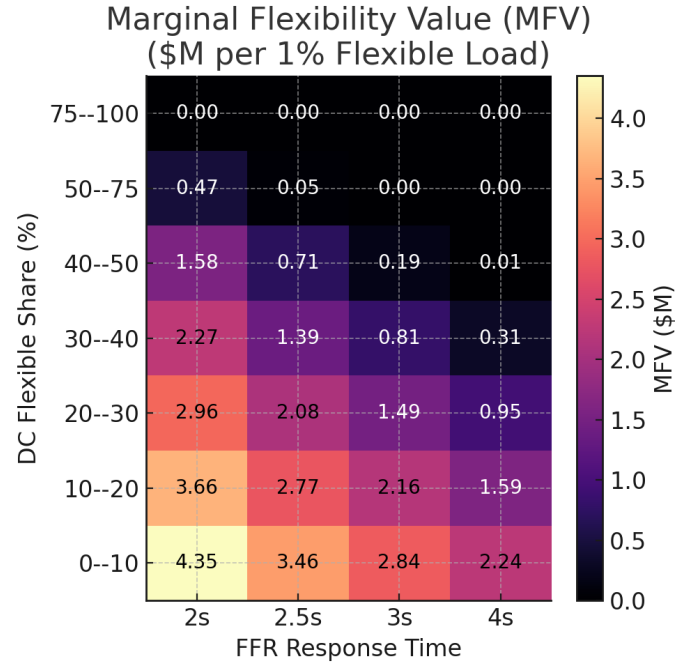


Fig. 8: Heatmap of MFV across different DC flexible load shares and FFR response times. Each cell denotes the additional cost saving (\$M) per 1% increase in flexible load share. Darker colors indicate higher marginal benefit.

capacity expansion trends. In this scenario, the total data center capacity is increased from 1000 MW to 2000 MW, representing a significantly larger share of the system load. To ensure consistency, system-wide demand and renewable generation capacity are also scaled proportionally, maintaining the original wind penetration ratio. This setup enables a fair comparison between present and future system configurations.

Table II presents the MFV values under current and future data center capacities, assuming a 2 s response delay. The results highlight two key trends. First, as the flexible share increases, the marginal benefit of additional flexibility gradually declines, reflecting saturation of the system’s frequency reserve requirements. Second, and more importantly, the total economic value of FFR from data centers increases substantially in the 2030 scenario. Early-stage flexibility (0–10%) yields more than twice the cost reduction per percentage point compared to the current scenario, indicating that as data centers become a larger portion of the system load, their fast-response capability becomes more impactful and valuable from a system-level perspective.

As shown in Table III, increasing the proportion of flexible load within data centers not only improves system economics, but also enhances the effective integration of wind energy. In both present (1000 MW) and future (2000 MW) data center capacity scenarios, the share of wind power in total generation grows with greater DC-based flexibility. This reflects the improved system capability to accommodate variable renewable generation without violating frequency security constraints.

TABLE II: MFV comparison for 1000 MW and 2000 MW DC capacities under 2 s response delay

DC Flexible Share (%)	DC Capacity	
	1000 MW	2000 MW
0–10	4.35	9.16
10–20	3.66	8.38
20–30	2.96	7.55
30–40	2.27	6.71
40–50	1.58	5.86
50–75	0.47	4.28
75–100	0.00	0.35

TABLE III: Wind power share under different DC flexible load ratios with 2s FFR delay in current and future capacity scenarios

DC Flexible Share (%)	Wind Power Share (%)	
	Benchmark scenario	2030 scenario
0–10	18.3	23.5
10–20	19.4	26.8
20–30	20.1	28.2
30–40	20.3	29.3
40–50	20.5	30.0
50–75	20.6	30.5
75–100	20.7	30.9

C. Methods Comparison

To further demonstrate the advantages of the proposed decision tree-based constraint learning (DT-CL) method, we conduct comparative studies against two benchmark linearization approaches: (1) conventional piecewise linear approximation (PLA), and (2) K-means-based regional linearization (KRL).

The PLA method discretizes nonlinear relationships into manually defined linear segments across input ranges, commonly used for embedding nonlinearity into MILP formulations. The KRL, on the other hand, employs unsupervised clustering to group historical samples into linearly separable regions, followed by local linear fitting within each cluster.

Table IV provides a quantitative comparison of the three methods applied to the Safe UC problem. All approaches result in an equal number of linear constraints, ensuring a fair comparison in terms of model complexity. The Feasibility Pass Rate represents the proportion of operating points in the validation set (8,760 hourly time steps) where the linearized constraints, when substituted back into the original non-linear (NL) constraint functions, are satisfied. The Average NL Constraint Error measures the mean relative deviation between the linearized and NL constraint values across all validation points, indicating the approximation error introduced by the linearization process. A lower value implies a closer match between the linearized constraints and the original NL representation.

Despite the parity in constraint count, DT-CL achieves the lowest system operation cost, reducing expenditure by 5.0% relative to PLA and 2.9% relative to KRL. This improvement

TABLE IV: Quantitative Comparison of Linearization Methods in the UC Model

Metric	PLA	K-means	DT-CL (Ours)
Linear Constraints	2,951,904	2,951,904	2,951,904
MILP Solving Time (s)	36.72	37.23	36.51
System Operation Cost (\$M)	2298.17	2249.32	2183.81
Feasibility Pass Rate (%)	97.2	98.6	100
Average NL Constraint Error	0.132	0.00561	0.00128

stems from its higher linearization fidelity: the Average NL Constraint Error of DT-CL (0.00128) is over two orders of magnitude lower than PLA (0.132) and more than four times lower than KRL (0.00561). The higher accuracy also translates into perfect physical feasibility, as evidenced by a 100% Feasibility Pass Rate, compared with 97.2% for PLA and 98.6% for KRL. In practical terms, this means that DT-CL produces UC schedules that satisfy the original nonlinear nadir frequency constraint in all hourly operating conditions, avoiding the security violations that occur in the benchmark methods.

The MILP solving times are comparable across all methods, with differences within 2%, indicating that the improved accuracy of DT-CL does not introduce additional computational burden. Taken together, these results demonstrate that DT-CL offers a superior trade-off between accuracy, feasibility, and tractability, making it an effective approach for embedding dynamic security constraints into large-scale UC formulations.

In addition to the linearization-based methods, we also experimented with solving the original nonlinear nadir constraint formulation using a gradient-based optimization approach based on the Adaptive Moment Estimation (ADAM) [23] algorithm—a first-order method widely used in deep learning. However, this approach exhibited prohibitively long convergence times and frequently failed to identify feasible solutions within acceptable tolerances. Due to these practical limitations, it is excluded from the main comparative analysis.

V. CONCLUSION

As data centers continue to expand in scale and energy intensity, their role in modern power systems is becoming increasingly critical. The growing share of data center demand poses challenges for system reliability, but also presents new opportunities for flexible load management. In particular, leveraging data center flexibility for frequency support can enhance system stability while reducing operational costs. However, effectively integrating this flexibility into power system scheduling requires tractable models that respect both physical limits and data center operational constraints.

To address this challenge, we propose a Safe UC framework that embeds data-driven frequency security constraints combining data center frequency response into a traditional UC model. The core innovation lies in a constraint learning module based on decision tree classification, trained on historical frequency response data. By translating the learned structure into piecewise linear constraints, the method produces inter-

pretable, MILP-compatible safety constraints that capture real-world system behavior under uncertainty.

We evaluate the system-level impact of flexible data center participation through two representative scenarios: a present-day benchmark and a projected 2030 future with doubled data center capacity. Results show that increasing the share of fast-responding flexible load within data centers leads to significant reductions in total system cost. However, this benefit is subject to diminishing returns, as revealed by the proposed Marginal Flexibility Value metric. For example, under a 2-second response delay, the MFV reaches \$4.35M per additional 1% of flexible load in our benchmark scenario, rising to \$9.16M in the 2030 scenario.

Beyond economic savings, enhanced data center flexibility also facilitates greater integration of renewable energy. In the 2030 case, the share of wind generation increases from 23.5% to 30.9% as DC flexibility expands from 0% to 100%, highlighting a strong co-benefit between frequency-secured scheduling and decarbonization goals.

In summary, the Safe UC framework offers a scalable and interpretable approach to co-optimizing economic dispatch, frequency security, and renewable integration. Future work will extend this framework to incorporate stochastic uncertainty, geographically distributed data center clusters, and intelligent workload scheduling for dynamic load shaping.

REFERENCES

- [1] International Energy Agency, "Energy and ai," <https://www.iea.org/reports/energy-and-ai>, International Energy Agency, 2025, iEA, Paris. Licence: CC BY 4.0.
- [2] "Powering intelligence: Analyzing artificial intelligence and data center energy consumption," Electric Power Research Institute (EPRI), Palo Alto, CA, White Paper 3002028905, Mar. 2024, available online: <https://www.epri.com/research/products/000000003002028905>.
- [3] Y. Zhou, A. Paredes, C. Essayeh, and T. Morstyn, "Ai-focused hpc data centers can provide more power grid flexibility and at lower cost," *arXiv preprint arXiv:2410.17435*, 2024.
- [4] J. Li, Z. Bao, and Z. Li, "Modeling demand response capability by internet data centers processing batch computing jobs," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 737–747, 2014.
- [5] Y. Zhang, D. C. Wilson, I. C. Paschalidis, and A. K. Coskun, "Hpc data center participation in demand response: An adaptive policy with qos assurance," *IEEE transactions on sustainable computing*, vol. 7, no. 1, pp. 157–171, 2021.
- [6] Z. Ding, S. Chen, Y. Sun, K. Shi, J. Wang, S. Chen, T. Xiao, Y. Wang, and X. Wei, "Data center job scheduling and energy management under uncertain environments," *IEEE Transactions on Industry Applications*, 2025.
- [7] S. Loganathan, R. D. Saravanan, and S. Mukherjee, "Energy aware resource management and job scheduling in cloud datacenter," *International Journal of Intelligent Engineering & Systems*, vol. 10, no. 4, 2017.
- [8] Y. Sun, Z. Ding, Y. Yan, Z. Wang, P. Dehghanian, and W.-J. Lee, "Privacy-preserving energy sharing among cloud service providers via collaborative job scheduling," *IEEE Transactions on Smart Grid*, 2024.
- [9] T. Yang, H. Jiang, Y. Hou, and Y. Geng, "Carbon management of multi-datacenter based on spatio-temporal task migration," *IEEE Transactions on Cloud Computing*, vol. 11, no. 1, pp. 1078–1090, 2021.
- [10] J. Zheng, A. A. Chien, and S. Suh, "Mitigating curtailment and carbon emissions through load migration between data centers," *Joule*, vol. 4, no. 10, pp. 2208–2222, 2020.
- [11] S. Chen, P. Li, H. Ji, H. Yu, J. Yan, J. Wu, and C. Wang, "Operational flexibility of active distribution networks with the potential from data centers," *Applied Energy*, vol. 293, p. 116935, 2021.
- [12] A. Z. G. Seyyedi, E. Akbari, S. M. Rashid, S. A. Nejati, and M. Gitzadeh, "Application of robust optimized spatiotemporal load management of data centers for renewable curtailment mitigation," *Renewable and Sustainable Energy Reviews*, vol. 204, p. 114793, 2024.
- [13] Z. Chu, U. Markovic, G. Hug, and F. Teng, "Towards optimal system scheduling with synthetic inertia provision from wind turbines," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4056–4066, 2020.
- [14] L. Meng, J. Zafar, S. K. Khadem, A. Collinson, K. C. Murchie, F. Coffele, and G. M. Burt, "Fast frequency response from energy storage systems—a review of grid standards, projects and technical issues," *IEEE transactions on smart grid*, vol. 11, no. 2, pp. 1566–1581, 2019.
- [15] F. Teng, V. Trovato, and G. Strbac, "Stochastic scheduling with inertia-dependent fast frequency response requirements," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1557–1566, 2015.
- [16] H. Li, Y. Qiao, Z. Lu, B. Zhang, and F. Teng, "Frequency-constrained stochastic planning towards a high renewable target considering frequency response support from wind power," *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 4632–4644, 2021.
- [17] R. Lahon, S. Stanley, C. O'Dwyer, M. Devine, and D. Flynn, "Impact of wide-scale data centre growth on power system operation with large share of renewables," in *2020 17th International Conference on the European Energy Market (EEM)*. IEEE, 2020, pp. 1–6.
- [18] B. Jiang, C. Guo, and Z. Chen, "Frequency constrained unit commitment considering reserve provision of wind power," *Applied Energy*, vol. 361, p. 122898, 2024.
- [19] H. Huang and F. Li, "Sensitivity analysis of load-damping characteristic in power system frequency regulation," *IEEE transactions on power systems*, vol. 28, no. 2, pp. 1324–1335, 2012.
- [20] H. Jia, Q. Hou, P. Yong, F. Teng, G. Strbac, C. Fang, and N. Zhang, "Learning multiple convex voltage stability constraints for unit commitment," *IEEE Transactions on Power Systems*, 2024.
- [21] R. Christie, "Power systems test case archive," *University of Washington, Department of Electrical Engineering*, vol. 20895, no. 98195, 1993, <https://labs.ece.uw.edu/pstca/>.
- [22] U. P. Networks, "Standard profiles uk power networks uses for electricity demand," <https://ukpowernetworks.opendatasoft.com/explore/dataset/ukpn-standard-profiles-electricity-demand/information/>, 2024, accessed: March 27, 2025.
- [23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. [Online]. Available: <https://arxiv.org/abs/1412.6980>