Uncertainty-aware Day-ahead Datacenter Workload Planning with Load-following Small Modular Reactors

YIJIE YANG, The Hong Kong Polytechnic University, Hong Kong
DAN WANG, The Hong Kong Polytechnic University, Hong Kong
JIAN SHI, University of Houston, USA
CHENYE WU, The Chinese University of Hong Kong - Shenzhen, China
ZHU HAN, University of Houston, USA

The rapid rise of AI applications has driven datacenters to unprecedented energy demands, which has prompted major tech companies to adopt on-site nuclear power plants (NPPs) alongside grid electricity. While existing research focuses on off-site NPPs in multi-energy systems optimized for investment returns, recent advances in small modular reactors (SMRs), particularly load-following SMRs (LF-SMRs), offer flexible, reliable power tailored for datacenter co-location. However, LF-SMRs are governed by a set of physical constraints, such as ramp rate and stability limits, making them unsuitable as fully dispatchable sources. This paper proposes a novel day-ahead workload scheduling approach that jointly coordinates datacenter operations and LF-SMR output, explicitly modeling these constraints. We develop a two-stage formulation that forecasts carbon-free grid energy from the grid using conformal prediction in the first stage and then optimizes LF-SMR output and workload scheduling via mixed-integer programming in the second stage. Evaluation on real workload traces shows that our method reduces carbon-based energy consumption by up to 43.44% compared to baselines that omit nuclear integration or ignore SMR limitations.

CCS Concepts: • **Social and professional topics → Sustainability**; • **Hardware → Enterprise level and data centers power issues**.

Additional Key Words and Phrases: Sustainable computing, Datacenter, Decarbonization, Nuclear energy, Carbon-aware scheduling

1 INTRODUCTION

We have seen a rapid growth of AI applications in the past decade [22]. To run these applications, significant computing resources are needed [28]. GPUs have become more powerful, and increasingly energy consuming [53]. As an example, the NVIDIA H100 SXM5 has a thermal designed power of 1100 watts, while the forthcoming next-generation B200 is expected to consume 2000 watts [33].

Datacenters are experiencing unprecedented energy demands driven by the rapid expansion of AI applications. In particular, the United Nations has advocated for a 24/7 carbon-free energy roadmap, calling for datacenters to be powered exclusively by carbon-free energy around the clock [15, 26, 48]. Among the available options, nuclear energy has emerged as a promising solution due to its carbon-free nature and improving safety record. For example, nuclear power is associated with just 0.07 deaths per terawatt-hour of electricity produced—significantly lower than lignite, which causes 32.72 deaths per terawatt-hour due to accidents and air pollution [16, 17]. In response, major technology companies have begun acquiring nuclear power resources to support their datacenter operations. Google has signed a clean nuclear energy agreement and plans to launch its first nuclear reactor by 2030 [20]; Meta issued a request for proposals in 2024 for up to 4 GW of new nuclear capacity to come online in the early 2030s [1]; and Microsoft has acquired and reopened the nuclear power plant at Three Mile Island [32].

Traditionally, nuclear power plants (NPPs) have been studied in the context of electricity markets and multi-energy system operation/dispatch, where they serve as stable, large-scale energy sources [31]. In multi-energy systems, NPPs can co-generate heat and electricity for district use [39, 41], while other studies explore NPPs in energy markets for providing electricity, reserve services, and thermal products [40, 47]. These works typically focus on investment-level planning and evaluate reactor capacities, technologies, and long-term returns on investment. For example, one recent study optimized multi-year NPP investment strategies for powering datacenters [44].

Advancements in nuclear technologies have led to the development of small modular reactors (SMRs), which offer enhanced safety, flexibility, and load-following capabilities. In particular, load-following SMRs (LF-SMRs) can adjust power output dynamically in response to real-time demand, making them ideal for co-location with datacenters [4, 11]. However, operating LF-SMRs alongside datacenters requires careful coordination at the operational level, as their dynamic response is constrained by physical limitations. Specifically, LF-SMRs cannot be treated as on-demand, immediately dispatchable energy sources due to the following two constraints: (1) the ramp-rate restriction, which limits the rate at which the reactor can change its power output, and (2) the stable power period restriction, which mandates that once the reactor reduces output, it must maintain a stable level for a minimum duration (typically 2 to 9 hours) before ramping back up [19]. Moreover, excess energy generation from the LF-SMR can result in inefficient curtailment penalties, as unused nuclear energy imposes grid stability and economic challenges [8]. These operational constraints unfold on an hourly timescale, aligning with the temporal dynamics of renewable energy availability and datacenter workload shifting in day-ahead planning.

To address the critical challenge of coordinating datacenter operations with on-site nuclear generation, this paper presents a novel approach for the day-ahead co-optimization of datacenter workloads and LF-SMR output. We focus on a representative scenario where a datacenter is connected to the bulk power grid, which provides

Authors' addresses: Yijie Yang, yi-jie.yang@connect.polyu.hk, The Hong Kong Polytechnic University, Hong Kong; Dan Wang, dan.wang@polyu.edu.hk, The Hong Kong Polytechnic University, Hong Kong; Jian Shi, jshi23@central.uh.edu, University of Houston, Houston, Texas, USA; Chenye Wu, chenyewu@yeah.net, The Chinese University of Hong Kong - Shenzhen, Shenzhen, China; Zhu Han, hanzhu22@gmail.com, University of Houston, Houston, Texas, USA.

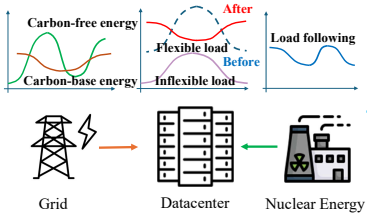

Fig. 1. The framework of DCSMR

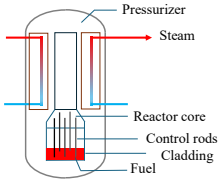

Fig. 2. An LF-SMR

a mix of carbon-free (e.g., solar, wind) and carbon-based (e.g., coal, oil) energy sources (see Fig. 1). To ensure a cleaner energy supply, datacenters often establish power purchase agreements (PPAs) that define the share and type of carbon-free electricity they receive from the grid [5, 12]. Within the datacenter, workloads are classified into inflexible tasks (e.g., real-time inference) and flexible tasks (e.g., model training), enabling intelligent scheduling decisions. By forecasting the availability of carbon-free renewable energy under the PPA, the datacenter constructs day-ahead workload plans that shift flexible workloads to periods with greater clean energy availability [3, 50]. This carbon-aware scheduling paradigm is already being practiced at scale by companies such as Google [14, 38].

Our proposed approach extends the state-of-the-art carbon-aware workload planning by explicitly incorporating the physical and operational constraints of LF-SMRs, enabling datacenters to better align their sustainability goals with the realities of nuclear energy generation. To achieve this, we formulate a **Day-ahead Datacenter Workload Planning with LF-SMR (DCSMR-Plan)** problem, which models both the ramp-rate and stable power period restrictions inherent to LF-SMR operation, as well as the stochastic nature of carbon-free energy availability from the grid. We develop a two-stage solution framework that integrates a conformal prediction-based optimization module for managing uncertainty and a mixed-integer programming module for operational decision-making. Our method guarantees compliance with a user-defined cap on the combined share of renewable and nuclear carbon-free energy with statistical confidence. This capability empowers datacenter operators to meet decarbonization targets with measurable assurance, offer carbon-free computation as a quantifiable service metric, and manage carbon credit purchases more strategically and cost-effectively.

We evaluate our model and algorithm using the Google datacenter trace [43]. The trace has the power utilization information for 57 server clusters (the so-called power domains) within Google datacenters in May 2019. The proposed model and algorithm are compared to DC-Plan, which does not account for on-site SMRs, and DCSMR-PA, which does not consider SMR restrictions. The results show that our DCSMR-Plan outperforms DC-Plan in reducing carbon-based energy by up to 43.33% and DCSMR-PA by 30.32%. Additionally, when compared to the deterministic methods that do not account for prediction errors in renewable energy, our algorithm shows an improvement of 8.83%.

2 BACKGROUND AND RELATED WORK

Background on Nuclear Energy Generation: Fig. 2 shows an LF-SMR. The pressurizer has a reactor core, where nuclear fission takes place. In the nuclear fission, neutrons collides with nuclear fuel, such as uranium-235. This operation produces heat and the heat generates steam, which then drives a steam turbine to produce electricity. LF-SMR operations have restrictions.

First, the change of energy generation, i.e., the *ramp rate* of the reactor, has a limit. In the reactor core, control rods regulate the rate of nuclear fission with nuclear fuel. Inserting or withdrawing a control rod results in a power decrease or increase. This can lead to an immediate decrease or increase in the fuel temperature. Consequently, the fuel pellets contract or expand. Rapid changes in the core power can impose substantial thermal and mechanical stresses on the fuel pellets and cladding, potentially causing fuel cracking and cladding failure [21, 30]. Therefore, the ramp rate of the core power change is limited to ensure that these stresses remain within the design tolerances of the fuel assemblies.

Second, when the LF-SMR starts decreasing its power generation, there is a *stable power period* before the power generation can increase. This is because the Xe-135 concentration is influenced by the decay chain of I-135 and Xe-135. After a reactor reduces its power generation, the rate of xenon buildup exceeds its decay, causing the Xe-135 concentration to increase for several hours and then gradually decline to a new equilibrium (the half-life of I-135 is 6.6 hours and Xe-135 is 9.2 hours) [45]. To address this, SMRs allow sufficient time for operators to implement reactivity changes to compensate for the xenon transient [34, 35].

Related Work: Taking SMRs as energy sources, recent studies have developed new models and optimizations on the electricity market [40], electricity-hydrogen integrated energy system [39] [18], electricity-heating multi-energy systems [41], etc. The primary objective is to minimize the energy generation costs of the power systems. Results show that SMRs can benefit power systems with greater revenue in providing electricity, thermal energy, and reserve services. One recent work on datacenters [44] studies appropriate SMRs for the optimal return of investment. It shows that, among various nuclear technologies, small modular reactors (SMRs) exhibit superior economic performance compared to large-scale nuclear power plants. Existing studies are less concerned on the operation level of SMRs. With on-site SMRs co-located with datacenters, new models and algorithms need to be developed.

Carbon-aware and energy-aware datacenters have attracted a lot of studies [2, 14]. There are studies on runtime optimization or day-ahead optimization. Runtime optimization minimizes the carbon by workload assignment [9, 10], frequency scaling [13], power capping [36], geographical server relocation [29], load shifting [2, 24], etc. These studies are orthogonal to this paper. Day-ahead optimization conducts workload planning. For example, Google classified inflexible workloads (e.g., LLM inference) and flexible workloads (e.g., AI training) and shifted the flexible workloads to the time period with abundant renewable energy [14, 38]. This paper supplements the studies in this thread with a new energy source from LF-SMR, and we develop new models and algorithms to coordinate the day-ahead operations of LF-SMRs and datacenters.

3 MODELS AND PROBLEM FORMULATION

This section presents the models that form the basis of our problem formulation. We develop the operational model and physical constraints of the LF-SMR in Section 3.1. The datacenter power supply model is described in Section 3.2. The datacenter power supplies come from the power grid. A datacenter usually signs a power purchase agreement (PPA) with the power grid. PPAs have many contract formats, see details [6, 49]. In this paper, we study a common format, where the power grid allocates the carbon-free renewable power (e.g., solar or wind) generated from a certain project to the datacenter and supplements the datacenter with carbon-based energy when there is a shortage.[1] Section 3.3 outlines the datacenter's workload-driven power demand. The various power-related costs are presented in Section 3.4. Building upon these models, we formulate the Day-ahead Datacenter Workload Planning with LF-SMR Problem in Section 3.5.

3.1 The LF-SMR Model

We now model the power output of an LF-SMR. The power output of an LF-SMR p_t in a period t depends on the nuclear reaction between the fuel (U-235) and the control rods [35]. There are the following states in an LF-SMR: (1) the ramp-up state and the ramp-down state: the control rods are withdrawn from or inserted into the core, causing the power output to increase or decrease; (2) the stable state: the control rods remain stable in the core, and the power output remains constant. There are maximum rates at which reactors can adjust their power output, i.e., a ramp-down limit $\overline{P}_{RD}$ and a ramp-up limit $\overline{P}_{RU}$. The binary state variables St_t, Up_t, and Rd_t represent the stable output, ramp-up, and ramp-down states, respectively, and are governed by the following constraints:

$$Rd_t + Up_t + St_t = 1, \quad \forall t, \tag{1}$$

$$Rd_t, Up_t, St_t \in \{0, 1\}, \quad \forall t. \tag{2}$$

The physical constraints of an LF-SMR can be modeled as follows:

(1) Ramping limits. The ramp-down and ramp-up limits of power output for an LF-SMR should satisfy Eqs. (3)-(4):

$$p_{t-1} - p_t \leqslant \overline{P}_{RD} \times Rd_t - \delta \times Up_t, \quad \forall t, \tag{3}$$

$$p_t - p_{t-1} \leqslant \overline{P}_{RU} \times Up_t - \delta \times Rd_t, \quad \forall t, \tag{4}$$

where δ is an auxiliary parameter (very small number, e.g., $1e10^{-4}$) used to force $St_t = 1$ when there is no ramping activity.

(2) Stable power periods. The constraints for stable power periods should satisfy Eq. (5):

$$(Up_t - Up_{t-1}) \times T_h \leqslant \sum_{tt=t-T_h}^{t-1} (St_t + Up_t), \quad \forall t, \tag{5}$$

where T_h denotes the minimum number of hours that the SMR must remain constrained at a stable output level.

In addition, the output of SMRs should satisfy the minimum and maximum output limits:

$$P_{MIN} \leqslant p_t \leqslant P_{MAX}, \quad \forall t, \tag{6}$$

where P_{MIN} and P_{MAX} represent the minimum and maximum output level of the SMR, respectively.

3.2 The Datacenter Power Supply Model

Datacenters source their electricity from the power grid, which provides a combination of energy types. At any time t, this supply can be categorized into: (1) carbon-based energy q_t: This portion is typically dispatchable, meaning its procurement can be determined by the datacenter; (2) carbon-free energy w_t: primarily sourced from intermittent renewables like solar and wind, this energy is generated externally and delivered to the datacenter.

There is uncertainty in carbon-free energy delivery [12]. [2] This stochastic nature is captured by modeling the actual carbon-free energy received, w_t, as:

$$w_t = (1 + \epsilon_t) W_t, \tag{7}$$

where W_t denotes the predicted value and ϵ represents the uncertainty associated with the prediction error.

The total power supply available to the datacenter from the grid, g_{grid} , is therefore the sum of these components:

$$g_{grid,t} = q_t + w_t. \tag{8}$$

3.3 The Datacenter Power Demand Model

Datacenter power demand is highly correlated to the workload. We first model the workloads, and then we derive the power demand from the workloads. Let the datacenter workloads be z_t. The workload can be divided into different classes according to their temporal flexibility. We follow the mode in [14]. Let jobs class be $c \in C$. $s_{c,k}$ is the aggregate load of class c jobs submitted at time k. Then the total workload is:

$$z_t = \sum\nolimits_{c \in C} \sum\nolimits_{k \in \mathcal{H}} Y_{k,c,t} \cdot s_{k,c}, \tag{9}$$

where $Y_{k,c,t} \geq 0$ denotes the fraction of the load $s_{c,k}$ allocated for processing at time t.

Each class $c \in C$ has a temporal flexibility parameter $h_c \in \mathbb{Z}_{\geq 0}$. It represents the maximum delay tolerable for jobs in that class. Inflexible workloads have $h_c = 0$. Then the workload planning problem is formulated as:

$$\sum\nolimits_{t \in \mathcal{T}} Y_{k,c,t} \geq 1, \quad \forall k \in \mathcal{H}, c \in C, \tag{10}$$

$$Y_{k,c,t} = 0, \quad \forall k \in \mathcal{H}, c \in C, t \notin \mathbb{Z}_{[k:k+h_c]}. \tag{11}$$

Eq. (10) ensures the full allocation of compute loads, while Eq. (11) enforces temporal flexibility by restricting scheduling to within the allowable delay window $[k, k + h_c]$.

The power consumption v_t at time t is modeled as follows:

$$v_t = e(z_t), \tag{12}$$

where $e(\cdot)$ represents the workload-to-energy consumption model, as described in [38] and [25].

[1] Carbon-free renewable energy generation exhibits significant hourly and seasonal supply fluctuations. In most cases, the contracted project with renewable energy cannot fully cover the demands, while over-subscription may result in inefficient curtailments, wherein renewable energy generation is deactivated to align supply with demand.

[2] There are contracts where the seller guarantees all electricity is renewable. Yet the electricity price is high and, from the perspective of "greenness", it is at the sacrifice of other buyers since non-renewable energy is used to cover the insufficiency between the demands and the renewable energy generation.

3.4 The Datacenter Power Cost Model

There are three types of costs: (1) the cost associated with the energy generation from the power grid, (2) the cost associated with the nuclear energy generation, and (3) the cost associated with the curtailment of nuclear energy of the LF-SMR.

Energy Purchased Cost. The energy purchased cost accounts for the expenses associated with procuring electricity and the carbon cost of utilizing carbon-based energy. It is expressed as:

$$EPC = \sum_{t=1}^{T} \alpha q_t + p_{co_2} q_t I_t + \kappa w_t, \tag{13}$$

where α and κ represent the price of carbon-based and carbon-free energy, respectively. p_{co_2} denotes the carbon tax. I_t is the carbon intensity of the energy from the grid.

Nuclear Generation Cost. The generation cost associated with the SMR is formulated as:

$$EGC = \sum_{t=1}^{T} \gamma p_t, \tag{14}$$

where p_t represents the nuclear energy generated, and γ denotes the generation cost of the SMR.

Energy Excess Cost. If the energy generated by the LF-SMR is greater than the demands, the excessive energy has to be handled [37], e.g., by storage, etc., and this brings about a penalty cost.

$$ECC = \sum_{t=1}^{T} \beta(p_t - d_t), \tag{15}$$

where d_t is the energy consumed by the datacenter. β denotes the excess cost per unit.

3.5 Problem Formulation

We now present the Day-ahead Datacenter Workload Planning with Load-following Small Modular Reactor Problem (DCSMR-Plan).

$$\min EPC + NGC + EEC \tag{16}$$

$$\text{s.t.} \quad (1) - (11), \tag{17}$$

$$p_t \geqslant d_t, \quad \forall t, \tag{18}$$

$$g_{grid,t} + d_t = v_t, \quad \forall t, \tag{19}$$

$$w_t + d_t \geqslant \eta v_t, \quad \forall t, \tag{20}$$

where η indicates the green factor, which lies within the range $[0, 1]$. Eq. (18) indicates that the consumption of nuclear power must exceed its generation. Eq. (19) specifies that the power supply should be equal to the demand. Eq. (20) defines the green energy coverage, namely, the consumption of carbon-free green energy should be greater than a certain proportion. In our formulation, we directly add the three costs: EPC, NGC, and ECC. Different scenarios can emphasize different components through a weighted sum. We leave these for future work.

4 METHODOLOGY

We need to develop an algorithm to output (1) a day-ahead plan for the datacenter on the shifting schedule of its flexible workloads v_t and (2) a plan for LF-SMR on the power generation p_t. A key challenge addressed by DCSMR-Plan is the inherent uncertainty in carbon-free renewable energy supplied by the power grid, which necessitates robust forecasting and planning. To this end, DCSMR-Plan adopts a two-stage, uncertainty-driven forecasting and optimization algorithm.

DCSMR-Plan Algorithm. The overall procedure comprises two stages: the upstream prediction of carbon-free energy availability and the downstream optimization of datacenter and SMR operations. In the first stage, conformal prediction (CP) [23] is utilized to construct prediction intervals for carbon-free energy generation. In the second stage, based on the prediction intervals obtained from CP, the problem is reformulated as a robust optimization problem with a traditional box uncertainty set [7]. This reformulated problem is subsequently transformed into a standard Mixed Integer Linear Programming (MILP) problem, which can be solved using commercial solvers such as CPLEX or Gurobi. The detailed procedure is presented in Algorithm 1.

Algorithm 1 Conformal Prediction-Based DCSMR-Plan

Input: Datacenter workload profile $s_{c,k}$, SMR operation parameters T_h, $\overline{P}_{RU}$, $\overline{P}_{RD}$, and historical renewable energy data $\mathcal{D} = \{X_{w_i}, Y_{w_i}\}_{i=1}^{n}$.

Output: The SMR generation plan p_t and the datacenter workload scheduling plan v_t.

`/* Stage 1: Renewable Energy Prediction */`

1: Train a prediction model for renewable energy generation using the historical dataset $\mathcal{D}$ and predict renewable energy generation for the planning horizon.

2: Prediction intervals $\hat{C}(X_{w_{t+1}}) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$ are constructed using Conformal CP to serve as the uncertainty set, ensuring coverage as specified in Eq. (21).

`/* Stage 2: Optimization of DCSMR Problem */`

3: Solve problem (16) - (20) , get the optimal SMR generation plan p_t and datacenter workload scheduling plan v_t.

Algorithm Analysis. CP provides a rigorous, distribution-free methodology for generating prediction sets $\hat{C}(X_{w_{t+1}})$ for future carbon-free energy delivery $Y_{w_{t+1}}$, conditioned on features $X_{w_{t+1}}$. It ensures individual coverage at a specified confidence level, as defined by:

$$\mathbb{P}\left(Y_{w_{t+1}} \in \hat{C}(X_{w_{t+1}})\right) \geq \epsilon, \forall t, \tag{21}$$

where ϵ represents the confidence level. This guarantees that the true future renewable delivery $Y_{w_{t+1}}$ will fall within the predicted interval $\hat{C}(X_{w_{t+1}})$ with a probability of at least ϵ.

Based on the definitions of CP and robust optimization, we can obtain the corollary as follows:

Corollary 4.1 (Green Energy Coverage Guarantee). *If w_t is obtained using conformal prediction to determine the interval $Y_{w_t} \in \hat{C}(X_{w_t})$, the green energy coverage constraint Eq. (20) can be guaranteed with the probability of ϵ, namely,*

$$\mathbb{P}\left(Y_{w_t} + d_t \geqslant \eta v_t\right) \geq \epsilon$$

Brief summary: The interval $\hat{C}(X_{w_t})$ for renewable delivery at time t defines the uncertainty set for the corresponding variable in the robust optimization formulation. For example, if the datacenter requires a green energy coverage of 80% ($\eta = 0.8$), by employing an interval $[lb_t, ub_t]$ with a 95% confidence level (ϵ=0.95) as the uncertainty set, the resulting operational strategy is guaranteed to

be feasible even in the worst-case renewable generation scenario within that interval. Therefore, the 80% coverage target is achieved with the probability of 95%. This ensures that the generation from the nuclear power plant can adjust its output and still meet the green energy requirements of the data center within the probabilistically guaranteed range, despite fluctuations in renewable delivery.

5 EVALUATION

5.1 Evaluation Setup and Methodology

The datacenter workloads. We evaluate our model and algorithm using Google datacenter trace [43]. The trace has the power utilization information for 57 server clusters within Google datacenters in May 2019. The workload is directly quantified based on power demand. Specifically, power data from a single cluster is utilized for comparative analysis. The workload is assumed to be executed on a cluster with a peak power capacity of 20 MW, derived from actual normalized power data obtained from the trace. For simplicity, we only consider two class of jobs, the inflexible job and the flexible job, whose maximum delay time h_c is set to 5 hours. The cost of carbon-based electricity obtained from the grid, including carbon cost, is assumed to be \$100/MWh since we suppose that all of the energy from the grid, except PPA, is coal-based. The generation cost for the SMR is assumed to be \$50/MWh, and the cost associated with excess energy is assumed to be \$30/MWh.

The SMR energy generation. For the nuclear power plant, it is assumed to be a small modular reactor (SMR) with a capacity of 20 MW. We follow the widely adopted setting in nuclear power plant control [19] [42], the ramp rate is set as 10%, and the hold time T_h is specified as 3 hours.

The power supply from the power grid. We follow the setting in [51], all energy procured from the main grid, excluding renewable PPAs, is coal-fired. Four years of hourly wind and solar power data were generated using historical weather data, as described in [46]. The dataset was divided into two groups, with 50% allocated for training and 50% for testing.

Baselines: We compare DCSMR-Plan with three baselines: *(1) DC-Plan*: The day-ahead planning without coordination with the co-located SMR. The maximum wind power output is 30 MW, while the maximum solar power output is 42 MW. *(2) DCSMR-PA:* The day-ahead planning in coordination with the co-located SMR, where the SMR operation is simply modeled as an optimal fixed output. The physics restrictions are not modeled (Physics-restriction Agnostics), and power from gird is utilized to meet any remaining demand. *(3) DCSMRU-Plan:* This day-ahead planning coordinates with the co-located SMR, incorporating its physical constraints and utilizing forecasted renewable energy for optimization, but it does not account for uncertainty in renewable energy forecasts.

Evaluation metrics: *(1) Green Energy Coverage (GEC) [2]:* proportion of green energy in total datacenter consumption. It is used to evaluate the greenness of the datacenter. *(2) Nuclear Energy Utilization (NEU):* proportion of nuclear energy consumed to total nuclear energy produced. It is used to evaluate the extent of nuclear power utilization in the datacenter.

Cases. Three workload flexibility scenarios were considered: NoFlex (0% flexible), LowFlex (10% flexible), and HighFlex (30% flexible). Similarly, three renewable penetration scenarios were defined: LowRP (14 MW solar capacity), MedRP (28 MW solar capacity), and HighRP (42 MW solar capacity). All scenarios were evaluated over a two-year period with an hourly time resolution. All of our data and codes are available at GitHub [52].

5.2 Evaluation results

5.2.1 Performance under different DC workload flexibility. Fig. 3 presents the GEC and NEU for different methods under varying levels of data center workload flexibility. In the NoFlex scenario, the GEC of DC-Plan is observed to be 0.6736, while DCSMR-PA achieves a GEC of 0.7408. In contrast, the proposed DCSMR-Plan achieves a significantly higher GEC of 0.9662, representing a 43.44% improvement in non-renewable energy savings compared to DC-Plan. Additionally, DCSMR-Plan reduces non-renewable energy consumption by 30.42% compared to DCSMR-PA. In the HighFlex scenario, DCSMR-Plan achieves a GEC of 0.9895, representing a 41.15% increase in GEC compared to DC-Plan and a 25.39% reduction in non-renewable energy consumption compared to DCSMR-PA. It is noted that increasing datacenter flexibility enhances the GEC and NEU of all methods to some extent. However, the proposed DCSMR-Plan consistently outperforms the other methods across different scenarios.

5.2.2 Performance under different renewable penetration. Figure 4 presents the GEC and NEU under varying levels of renewable energy penetration. In scenarios with low renewable penetration, the GEC of DCSMR-Plan is 0.9960, while that of DCSMR-PA is 0.7707, representing an increase of 29.23%. Under conditions of high renewable penetration, the GEC of DCSMR-Plan decreases to 0.9320, whereas the GEC of DCSMR-PA drops to 0.6263, resulting in a larger difference of 48.81%. With respect to NEU, under scenarios of low renewable penetration, the NEU of DCSMR-Plan is 0.9276, compared to 0.9412 for DCSMR-PA. Under high renewable penetration conditions, the NEU of DCSMR-Plan decreases to 0.8875, resulting in a difference of 4.32%. In contrast, the NEU of DCSMR-PA declines to 0.6263, leading to a larger difference of 24.18%. This outcome is attributed to the increased power variability associated with higher levels of renewable energy penetration, which amplifies the limitations of DCSMR-PA in adjusting its output to meet demand. These results highlight the feasibility and robustness of the DCSMR-Plan in sustaining high levels of GEC and NEU across varying levels of renewable energy penetration.

5.2.3 Impacts of uncertainty. Fig. 5 illustrates the GEC and NEU under varying confidence levels (ϵ) of chance constraints, as well as the scheduling scenario that does not account for prediction uncertainty (DCSMRU-Plan). The results indicate that, under HighRP conditions, the GEC of DCSMRU-Plan is 0.8486, whereas the GEC of DCSMR-Plan with $\epsilon = 0.8$ increases to 0.9320, representing an improvement of 8.83%. Under LowRP conditions, the improvement is 2.12%. This difference arises because higher renewable penetration amplifies the impact of renewable energy prediction errors. Although increasing ϵ reduces NEU due to the need for robustness,

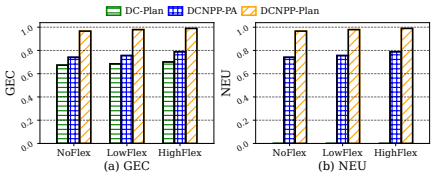

Fig. 3. Performance under various workload flexibility

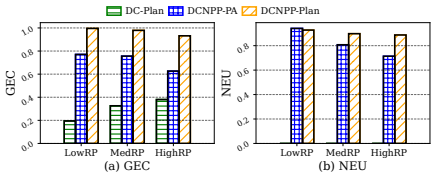

Fig. 4. Performance under various renewable penetration

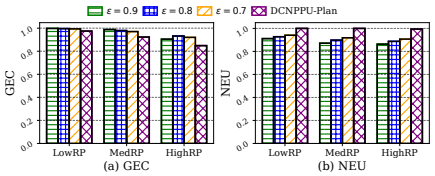

Fig. 5. Performance of DCSMR-Plan under different confidence levels ϵ

a trade-off must be made between ensuring green energy coverage and optimizing nuclear energy utilization.

6 CONCLUSION AND FUTURE WORK

This paper presents DCSMR-Plan, a novel day-ahead scheduling framework that co-optimizes datacenter workloads and LF-SMR operations while accounting for both nuclear operational constraints and renewable energy uncertainty. Through extensive evaluation using real-world datacenter traces and simulated energy scenarios, we demonstrate that DCSMR-Plan consistently outperforms baseline methods across varying levels of workload flexibility and renewable penetration. Key findings show that flexible workload scheduling, paired with coordinated LF-SMR control, can significantly enhance green energy coverage and nuclear energy utilization. Furthermore, the proposed DCSMR-Plan remains robust under high variability in renewable supply, offering a practical and scalable pathway for datacenters to meet 24/7 carbon-free energy goals while leveraging on-site nuclear power.

Future work can extend this foundation in several directions. One key area involves enhancing the modeling of physical system constraints and interactions with the main grid. This includes incorporating datacenter power ramp rate constraints to better reflect operational limitations and mitigate the impact of rapid load changes on both performance and grid stability [27]. Furthermore, extending beyond the simplified grid interaction employed here, integration of comprehensive power grid network models is warranted to capture realistic complexities such as transmission constraints and nodal dynamics. A second aspect for future work is the integration of advanced energy technologies. For example, LF-SMRs can co-generate thermal energy and, in next-generation designs, even hydrogen, introducing tightly coupled multi-output dynamics that require new modeling approaches. On the datacenter side, incorporating energy storage systems (ESS) could mitigate curtailment and enhance resilience. Finally, from a methodological standpoint, developing a multi-timescale optimization framework may improve planning accuracy and adaptability under forecast uncertainty.

ACKNOWLEDGMENTS

This work is in part supported by RGC GRF 15200321, 15201322, 15230624, ITC ITF-ITS/056/22MX, ITS/052/23MX, and PolyU 1-CDKK, G-SAC8. This work is also partially supported by NSF ECCS-2338158, ECCS-2302469, CMMI-2222810, Toyota. Amazon and Japan Science and Technology Agency (JST) Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE) JPMJAP2326.

REFERENCES

[1] Timothy Gardner. 2024. *Meta seeks nuclear power developers for reactors to start in early 2030s.* https://www.reuters.com/business/energy/meta-seeks-nuclear-power-developers-reactors-start-early-2030s-2024-12-03/
[2] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. Carbon explorer: A holistic framework for designing carbon aware datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 118–132.
[3] Mehdi Ahmadi, Lukas Knorr, and Henning Meschede. 2025. Improvement of Wind Power Utilization Through Flexible Operation of Data Center in Wind Parks. *Renewable Energy* (2025), 123073.
[4] Saeed Alhadhrami, Gabriel J Soto, and Ben Lindley. 2023. Dispatch analysis of flexible power operation with multi-unit small modular reactors. *Energy* 280 (2023), 128107.
[5] José Arellano and Miguel Carrión. 2023. Electricity procurement of large consumers considering power-purchase agreements. *Energy Reports* 9 (2023), 5384–5396.
[6] Cigdem Aslan and Tim Irwin. 2020. Power to the Fiscal? An Exploration of the Use of Credit Ratings to Estimate the Expected Cost of a Guarantee of a Power-Purchase Agreement. *An Exploration of the Use of Credit Ratings to Estimate the Expected Cost of a Guarantee of a Power-Purchase Agreement (June 5, 2020). World Bank Policy Research Working Paper* 9271 (2020).
[7] Aharon Ben-Tal, Arkadi Nemirovski, and Laurent El Ghaoui. 2009. Robust optimization. (2009).
[8] Lori Bird, Debra Lew, Michael Milligan, E Maria Carlini, Ana Estanqueiro, Damian Flynn, Emilio Gomez-Lazaro, Hannele Holttinen, Nickie Menemenlis, Antje Orths, et al. 2016. Wind and solar energy curtailment: A review of international experience. *Renewable and Sustainable Energy Reviews* 65 (2016), 577–586.
[9] Roozbeh Bostandoost, Walid A Hanafy, Adam Lechowicz, Noman Bashir, Prashant Shenoy, and Mohammad Hajiesmaili. 2024. Data-driven Algorithm Selection for Carbon-Aware Scheduling. *ACM SIGENERGY Energy Informatics Review* 4, 5 (2024), 148–153.
[10] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proceedings of the 2nd workshop on sustainable computer systems*. 1–7.
[11] Robin Gaster. 2025. *Small Modular Reactors: A Realist Approach to the Future of Nuclear Power.* Technical Report. Information Technology and Innovation Foundation.
[12] Yashar Ghiassi-Farrokhfal, Wolfgang Ketter, and John Collins. 2021. Making green power purchase agreements more predictable and reliable for companies. *Decision Support Systems* 144 (2021), 113514.
[13] Diandian Gu, Yihao Zhao, Peng Sun, Xin Jin, and Xuanzhe Liu. 2024. GreenFlow: A Carbon-Efficient Scheduler for Deep Learning Workloads. *IEEE Transactions on Parallel and Distributed Systems* (2024).
[14] Sophie Hall, Francesco Micheli, Giuseppe Belgioioso, Ana Radovanović, and Florian Dörfler. 2024. Carbon-Aware Computing for Data Centers with Probabilistic Performance Guarantees. *arXiv preprint arXiv:2410.21510* (2024).
[15] Savannah Goodman Hallie Cramer. 2024. *Accelerating a carbon-free future with hourly energy tracking.*
[16] Helmuth Boeck. 2022. *Why nuclear power is safer than ever.*
[17] Helmuth Boeck. 2025. *Safety of Nuclear Power Reactors.*
[18] Roshni Anna Jacob and Jie Zhang. 2023. Modeling and control of nuclear–renewable integrated energy systems: Dynamic system model for green electricity and hydrogen production. *Journal of Renewable and Sustainable Energy* 15, 4 (2023).
[19] Jesse D Jenkins, Zhi Zhou, Roberto Ponciroli, Richard B Vilim, Francesco Ganda, Fernando de Sisternes, and Audun Botterud. 2018. The benefits of nuclear flexibility in power system operations with renewable energy. *Applied energy* 222 (2018), 872–884.

[20] Joao da Silva . 2024. *Google turns to nuclear to power AI data centres.* https://www.bbc.com/news/articles/c748gn94k95o

[21] Jan Horst Keppler and Marco Cometto. 2012. *Nuclear energy and renewables: system effects in low-carbon electricity systems.* Technical Report.

[22] Laure de Roucy-Rochegonde and Adrien Buffard. 2025. *AI, Data Centers and Energy Demand: Reassessing and Exploring the Trends.*

[23] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1094–1111.

[24] Amy Li, Sihang Liu, and Yi Ding. 2024. Uncertainty-aware decarbonization for datacenters. *ACM SIGENERGY Energy Informatics Review* 4, 5 (2024), 141–147.

[25] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. 2024. Towards environmentally equitable AI via geographical load balancing. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems.* 291–307.

[26] Yuzhuo Li, Mariam Mughees, Yize Chen, and Yunwei Ryan Li. 2024. The unseen AI disruptions for power grids: LLM-induced transients. *arXiv preprint arXiv:2409.11416* (2024).

[27] Liuzixuan Lin and Andrew A Chien. 2023. Adapting datacenter capacity for greener datacenters and grid. In *Proceedings of the 14th ACM International Conference on Future Energy Systems.* 200–213.

[28] Liuzixuan Lin, Rajini Wijayawardana, Varsha Rao, Hai Nguyen, Emmanuel Wedan GNIBGA, and Andrew A Chien. 2024. Exploding ai power use: an opportunity to rethink grid planning and management. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems.* 434–441.

[29] Yejia Liu, Pengfei Li, Daniel Wong, and Shaolei Ren. 2024. Geographical Server Relocation: Opportunities and Challenges. *ACM SIGENERGY Energy Informatics Review* 4, 5 (2024), 34–43.

[30] Alexy Lokhov et al. 2011. Load-following with nuclear power plants. *NEA news* 29, 2 (2011), 18–20.

[31] D Michaelson and J Jiang. 2021. Review of integration of small modular reactors in renewable energy microgrids. *Renewable and Sustainable Energy Reviews* 152 (2021), 111638.

[32] Natalie Sherman. 2024. *Microsoft chooses infamous nuclear site for AI power.* https://www.bbc.com/news/articles/cx25v2d7zexo

[33] NVIDIA. 2024. *NVIDIA DGX B200,The foundation for your AI factory.* https://www.bbc.com/news/articles/cx25v2d7zexo

[34] OECD. 2021. *Technical and economic aspects of load following with nuclear power plants.* OECD Publishing.

[35] Roberto Ponciroli, Y Wang, Zhi Zhou, Audun Botterud, J Jenkins, RB Vilim, and FJNT Ganda. 2017. Profitability evaluation of load-following nuclear units with physics-induced operational constraints. *Nuclear Technology* 200, 3 (2017), 189–207.

[36] Haoran Qiu, Linghao Zhang, Chen Wang2 Hubertus Franke, Zbigniew T Kalbarczyk, and Ravishankar K Iyer. 2023. PARM: Adaptive resource allocation for datacenter power capping. In *Machine Learning for Systems Workshop at the Annual Conference on Neural Information Processing Systems (NeurIPS 2023).*

[37] Mohammad Amin Vaziri Rad, Alibakhsh Kasaeian, Xiaofeng Niu, Kai Zhang, and Omid Mahian. 2023. Excess electricity problem in off-grid hybrid renewable energy systems: A comprehensive review from challenges to prevalent solutions. *Renewable Energy* 212 (2023), 538–560.

[38] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, et al. 2022. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems* 38, 2 (2022), 1270–1280.

[39] Jubeyer Rahman, Roshni Anna Jacob, and Jie Zhang. 2025. Multi-timescale power system operations for electrolytic hydrogen generation in integrated nuclear-renewable energy systems. *Applied Energy* 377 (2025), 124346.

[40] Jubeyer Rahman and Jie Zhang. 2021. Optimization of nuclear-renewable hybrid energy system operation in forward electricity market. In *2021 IEEE Green Technologies Conference (GreenTech).* IEEE, 462–468.

[41] Jubeyer Rahman and Jie Zhang. 2023. Multi-timescale operations of nuclear-renewable hybrid energy systems for reserve and thermal product provision. *Journal of Renewable and Sustainable Energy* 15, 2 (2023).

[42] Jubeyer Rahman and Jie Zhang. 2024. Steady-state modeling of small modular reactors for multi-timescale power system operations with temporally coupled sub-models. *IEEE Transactions on Power Systems* (2024).

[43] Varun Sakalkar, Vasileios Kontorinis, David Landhuis, Shaohong Li, Darren De Ronde, Thomas Blooming, Anand Ramesh, James Kennedy, Christopher Malone, Jimmy Clidaras, et al. 2020. Data center power oversubscription with a medium voltage power plane and priority-aware capping. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems.* 497–511.

[44] Gabriel Jose Soto Gonzalez, Botros Naseif Hanna Bishara Hanna, Jakub Toman, Nahuel Guaita, Paul W Talbot, Christopher Shawn Lohse, and Aaron S Epiney. 2024. *Powering Data Centers with Clean Energy: A Techno-Economic Case Study of Nuclear and Renewable Energy Dependability.* Technical Report. Idaho National Laboratory (INL), Idaho Falls, ID (United States).

[45] Weston M Stacey. 2018. *Nuclear reactor physics.* John Wiley & Sons.

[46] Iain Staffell and Stefan Pfenninger. 2016. Using bias-corrected reanalysis to simulate current and future wind power output. *Energy* 114 (2016), 1224–1239.

[47] Shaojie Tan, Songbai Cheng, Kai Wang, Xiaoxing Liu, Hui Cheng, and Jun Wang. 2023. The development of micro and small modular reactor in the future energy market. *Frontiers in Energy Research* 11 (2023), 1149127.

[48] The United Nations. 2025. *24/7 Carbon-free Energy Compact.*

[49] Patrick Wallace. 2019. Long-term power purchase agreements: The factors that influence contract design. In *Research Handbook on International and Comparative Sale of Goods Law.* Edward Elgar Publishing, 305–333.

[50] Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 2021. Let's wait awhile: How temporal workload shifting can reduce carbon emissions in the cloud. In *Proceedings of the 22nd International Middleware Conference.* 260–272.

[51] Min Xiao and Ghassan Fadhil Smaisim. 2022. Joint chance-constrained multi-objective optimal function of multi-energy microgrid containing energy storages and carbon recycling system. *Journal of Energy Storage* 55, 11 (November 2022), 105842.

[52] Yijie Yang. 2025. *Github, DCSMR-Plan.* https://github.com/itsmeyijie/DCSMR-Plan

[53] Zhisheng Ye, Wei Gao, Qinghao Hu, Peng Sun, Xiaolin Wang, Yingwei Luo, Tianwei Zhang, and Yonggang Wen. 2024. Deep learning workload scheduling in gpu datacenters: A survey. *Comput. Surveys* 56, 6 (2024), 1–38.