DATA-DRIVEN LOW-CARBON OPERATION OF ENERGY SYSTEMS UNDER UNCERTAINTY

YIJIE YANG

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University
Department of Computing

Data-driven Low-carbon Operation of Energy Systems Under Uncertainty

Yijie Yang

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
July 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: ____________________

Name of Student: Yijie Yang

Abstract

Energy systems produce a significant share of greenhouse gas (GHG) emissions. To mitigate the impacts of climate change, it is imperative to curtail the carbon dioxide emissions originating from the energy sector. Despite the insights offered by existing research efforts, two important gaps remained. The first gap lies in the lack of incorporation of a carbon perspective in the operation of energy systems. Previous optimization problems are mainly cost-aware. New carbon-oriented operation strategies and algorithms need to be developed that allow decision-makers to define and plan the necessary short-term actions toward the goal of achieving zero-carbon, as well as to track progress over time. And this needs to coordinate different components and precisely calculate the carbon emissions within the energy systems. Another challenge lies in the uncertainties inherent in predictions. Downstream operational optimization problems, such as those relevant to energy systems, often require upstream predictions—for example, day-ahead forecasts for renewable energy generation, electricity load, electricity price, and carbon intensity of the main grid—as inputs. In this paper, we investigate carbon-aware operation scheduling for multi-energy building integrated energy systems (BIES), thermostatically controlled loads (TCLs), and data center energy systems (DCES).

Firstly, we investigate carbon-aware scheduling for multi-energy BIES. We explore achieving net-zero emissions through the carbon-aware optimal scheduling of the multi-energy BIES. We integrate advanced technologies and strategies, such as the carbon capture system (CCS), power-to-gas (P2G), carbon tracking, and emission allowance trading, into the traditional BIES scheduling problem. The proposed model enables accurate accounting of carbon emissions associated with building energy systems and facilitates the implementation of low-carbon operations. Furthermore, to address the challenge of accurately assessing uncertainty sets related to forecasting errors of loads, generation, and carbon intensity, we develop a learning-based robust optimization approach for BIES that is robust in the presence of uncertainty and guarantees statistical feasibility. The proposed approach comprises a shape learning stage and a shape calibration stage to generate an optimal uncertainty set that ensures favorable results from a statistical perspective. Numerical studies conducted based on both synthetic and real-world datasets have demonstrated that the approach yields up to 8.2% cost reduction, compared with conventional methods, in assisting buildings to robustly reach net-zero emissions.

Secondly, we focus on the carbon reduction potential of TCLs, which include air conditioners, heat pumps, water heaters, and refrigerators, play a pivotal role in demand response due to their thermal inertia and inherent flexibility. TCLs also substantially impact energy consumption and emissions within commercial and residential buildings, which makes them critical for the low-/zero-carbon transition that the building sector is undergoing to meet global climate objectives. To aid in this process, this paper proposes a carbon-aware robust scheduling approach for TCLs. The proposed approach precisely models carbon emissions attributed to TCLs, and formulates TCL scheduling as a distributionally robust chance-constrained (DRCC) optimization problem to ensure robust decision-making. We then develop a novel bilevel optimization reformulation strategy to address challenges such as over-conservatism and computational intractability that often arise from solving DRCC problems using conventional approaches. Real-world data evaluation demonstrates significant reductions in costs and carbon emissions compared to state-of-the-art methods, showcasing the effectiveness of our approach in potentially decarbonizing the building sector.

Third, we explore the carbon-aware scenario in data centers. The rapid growth of AI applications is dramatically increasing data center energy demand, exacerbating carbon emissions, and necessitating a shift towards 24/7 carbon-free energy (CFE). Unlike traditional annual energy matching, 24/7 CFE requires matching real-time electricity consumption with clean energy generation every hour, presenting significant challenges due to the inherent variability and forecasting errors of renewable energy sources. Traditional robust and data-driven optimization methods often fail to leverage the features of the prediction model (also known as contextual or covariate information) when constructing the uncertainty set, leading to overly conservative operational decisions. We propose a comprehensive approach for 24/7 CFE data center operation scheduling, focusing on robust decision-making under renewable generation uncertainty. This framework leverages covariate information through a multi-variable conformal prediction (CP) technique to construct statistically valid and adaptive uncertainty sets for renewable forecasts. The uncertainty sets directly inform the chance-constrained programming (CCP) problem, ensuring that chance constraints are met with a specified probability. We further establish theoretical underpinnings connecting the CP-generated uncertainty sets to the statistical feasibility guarantees of the CCP. Numerical results highlight the benefits of this covariate-aware approach, demonstrating up to 6.65% cost reduction and 6.96% decrease in carbon-based energy usage compared to conventional covariate-independent methods, thereby enabling data centers to progress toward 24/7 CEF.

Finally, we focused on nuclear-powered data center energy systems. The rapid rise of AI applications has driven data centers 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. We propose 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.

In summary, we propose four methods to overcome the challenges brought by the uncertainty of predictions in carbon-aware scheduling of different energy systems. All these methods are implemented and evaluated with real-world datasets, and the experimental results present the effectiveness of the proposed methods.

Publications Arising from the Thesis

1. Yijie Yang, Jian Shi, Dan Wang, Chenye Wu, and Zhu Han, “Carbon-Aware Scheduling of Thermostatically Controlled Loads: A Bilevel DRCC Approach”, in *IEEE Transactions on Smart Grid (TSG)* (2025).

2. Yijie Yang, Jian Shi, Dan Wang, Chenye Wu, and Zhu Han, “Net-Zero Scheduling of Multi-energy Building Energy Systems: A Learning-based Robust Optimization Approach with Statistical Guarantees”, in *IEEE Transactions on Sustainable Energy (TSTE)* (2024).

3. Yijie Yang, Dan Wang, Jian Shi, Chenye Wu, and Zhu Han, “Uncertainty-aware Day-ahead Datacenter Workload Planning with Load-following Small Modular Reactors”, accepted by *HotCarbon Workshop on Sustainable Computer Systems 2025*.

4. Yijie Yang, Jian Shi, Dan Wang, Chenye Wu, and Zhu Han, “A Conformal Prediction-Based Chance-Constrained Programming Approach for 24/7 Carbon-Free Data Center Operation Scheduling”, manuscript submitted to *IEEE Transactions on Smart Grid (TSG)*.

5. Xiaoyang Zhang, Yijie Yang, and Dan Wang, “Spatial-Temporal Embodied Carbon Models with Dual Carbon Attribution for Embodied Carbon Account-

ing of Computer Systems", manuscript submitted to *IEEE Transactions on Computers (TC).*

6. Xiaoyang Zhang, Yijie Yang, Dan Wang, "Spatial-temporal embodied carbon models for the embodied carbon accounting of computer systems", in *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems (E-Energy)*, pp. 464-471, 2024.

7. Jian Shi, Zirui Tong, Dan Wang, Chenye Wu, Zhu Han, Yijie Yang, and Kaushik Rajashekara, "A Multi-Layered Framework for Advancing Rapid and Balanced Electric Power System Decarbonization", manuscript submitted to *PNAS Nexus*.

8. Zirui Tong, Jian Shi, Yijie Yang, Dan Wang, Chenye Wu, Zhu Han, "Co-Optimizing Carbon Allowance Trading and Power Dispatch With Spot–Futures Market Coupling Under Uncertainty", manuscript submitted to *Energy and Climate Change*.

Acknowledgments

After I came to Hong Kong, I discovered a passion for hiking. While on MacLehose Trail Sec. 2, I was utterly exhausted by the repeated uphill and downhill. It was then that I suddenly remembered a verse from a fortune-telling paper I used to draw when I felt upset, which said: "After traveling thousands of miles through the mountains, don't complain that this road is full of obstacles." Then I realized that this verse can also describe my whole Ph.D. time. It may not be so accurate, as the end of the Ph.D. is just the beginning of one's research path.

First, I want to thank my supervisor, Prof. Dan Wang, whose passion and rigour for research impressed and inspired me. He always provided insightful suggestions and critical feedback, and he also cultivated in me a taste for promising research. Then I am deeply indebted to Prof. Jian Shi and Prof. Zhu Han at UH, and Prof. Chenye Wu at CUHK-SZ. They taught me how to look for research ideas and how to effectively present the work. Most importantly, they imparted the crucial lesson that enjoying the process and finding happiness in research are the most vital.

I made some good friends in HK. Dr. Shiping Shi, who offered support both, literally giving me a hand on Sharp Peak, known as the most dangerous trail in HK, and during my low ebb; who asked me if I wanted to have lunch together at the beginning of our friendship; who gave me a shot on the tennis court under the hot summer sun, shines bright like a star on winter nights in my life. I am also grateful to Dr. Fang He and Dr. Rui Lu. We explored many places in Hong Kong together, and they provided significant advice and help throughout my Ph.D. journey. Thanks also to Mr. Tao Wang, who provided much support and many snacks. My thanks extend to Dr. Shuhui Sun, Miss Wanchun Ye, Miss Ran Yue, Mr. Xiaoyang Zhang, and Dr. Yang Deng for their help and support. Finally, my deepest appreciation goes to my best friend, Chuhan Zhang, for her companionship, although we are friends who do not need to say thanks to each other.

Last but not least, I would like to express my heartfelt appreciation to my family, my parents and grandparents, who always gave me much comfort and support in my life. This was the very first page, not where the story ends.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

To mitigate the impacts of climate change and limit global temperature rise to 1.5°C above pre-industrial levels, as mandated by the Paris Agreement, a dramatic reduction in carbon emissions from the energy sector is essential. Meeting this critical goal requires global carbon emissions to decline by at least 45% by 2030 and reach net-zero by 2050 to avert catastrophic climate consequences [23]. However, the current trajectory is insufficient. In 2023, global energy-related carbon emissions paradoxically grew by 1.1%, increasing by 410 million tonnes to a new record high of 37.4 billion tonnes [10]. This alarming trend underscores the urgent need for accelerated and effective decarbonization efforts.

The energy system is the dominant source of global emissions, accounting for approximately 75.7% of overall greenhouse gas (GHG) emissions in 2021, with fossil fuels being the primary contributor [24]. Within this sector, coal-fired power generation remains a significant challenge, responsible for 30% of all energy-related carbon dioxide emissions [21].

Transitioning to low-carbon energy systems is essential for reducing greenhouse gas emissions. This transition is supported by many technologies and strategies. A cornerstone approach is shifting energy generation to low-carbon sources like wind, solar, and nuclear power. In 2024, renewable energy represented 92.5% of the total global power capacity expansion, adding 4,448 GW and reflecting robust annual growth of 15.1% [29]. Complementary technologies are also critical. Carbon capture and storage (CCS) systems, which typically involve three stages: capture, transport, and storage, help manage emissions, particularly from harder-to-abate sectors. In 2023, global CCS projects stored nearly 45 million tons of CO_2 annually [7]. Energy storage systems (ESS) are necessary to manage the intermittency of renewable energy sources, storing surplus energy for later discharge. Besides, demand response technologies manage energy demand, utilizing flexible loads such as thermostatically controlled loads (TCLs). Due to their unique thermal properties, TCLs can adjust power consumption temporarily with minimal comfort impact, allowing demand to be shifted to periods with sufficient renewable energy or low grid carbon intensity. Therefore, rapid deployment of low-carbon energy sources, coupled with essential supporting technologies like energy efficiency improvements, CCS, ESS, and demand response, are crucial components of a comprehensive transition strategy.

Despite this progress, realizing deep decarbonization within the energy system faces significant challenges. A key research challenge involves addressing the complex interplay between on-site energy technologies and the consumption of electricity from the utility grid. Specifically, achieving synergistic and carbon-aware coordination between diverse on-site assets, such as renewable energy generation and carbon capture systems, and external grid power is crucial. Energy systems often integrate these on-site components to reduce carbon footprints. However, the effective orchestration of these assets in conjunction with grid usage presents difficulties, particularly given that the increasing penetration of renewables leads to considerable temporal variability in grid carbon intensity. Consequently, developing strategies to schedule energy-intensive processes and manage flexible loads based on fluctuating carbon signals is critical for minimizing the system's total carbon emissions.

The second major challenge in low-carbon energy systems lies in managing the uncertainty inherent in forecasts. These include predictions for future renewable energy availability, electricity demand, electricity price, and grid carbon intensity. Renewable generation, for instance, is highly sensitive to weather patterns and diurnal variations, making accurate prediction difficult. While electricity demand exhibits predictable seasonal and daily patterns, it also contains inherent variability. Although forecasting models—typically based on historical data and machine learning techniques—can provide estimates [143], they inevitably introduce errors, especially under rapidly changing environmental conditions [191]. These prediction errors pose a significant threat to the reliability of emission-aware scheduling strategies, as misalignment between expected and actual clean energy availability can result in missed decarbonization opportunities or even system instability.

This thesis focuses on improving the robustness of the low-carbon operation of energy systems under uncertainty. The remainder of this chapter is structured as follows: Section 1.2 provides a brief introduction to the specific problems addressed in this research. Section 1.3 discusses the research framework adopted. Section 1.4 summarizes the main contributions of the thesis. Finally, Section 1.5 outlines the overall organization of the thesis.

1.2 The Problems

Deploying the uncertainty of prediction errors in energy systems incurs a lot of problems, and a wide range of research studies is required to overcome them. This thesis investigates methods for improving the robustness and carbon-awareness of energy system operations under uncertainty by addressing three specific problems:

1. Firstly, we investigate the carbon-aware optimal scheduling of multi-energy building integrated energy systems to achieve net-zero emissions. This work integrates various on-site technologies (e.g., carbon capture, renewable generation) and considers grid interaction based on fluctuating carbon intensity. A key focus is addressing uncertainty in forecast errors, such as loads, renewable generation, and carbon intensity predictions.

2. Secondly, we explore the carbon reduction potential of thermostatically controlled loads, recognizing their significant role in demand response and building decarbonization. A key challenge arises from uncertainties associated with the forecast errors of the building baseline load and outdoor temperatures.

3. Third, we address carbon-aware operation in data centers, specifically focusing on the transition to 24/7 carbon-free energy under increasing energy demand and renewable variability. The challenge is to robustly schedule data center operations under inherent renewable generation uncertainty while leveraging contextual information often ignored by traditional methods.

4. Finally, we explore the problem of coordinating data center operations with on-site nuclear generation, specifically load-following small modular reactors (LF-SMRs). The objective was to achieve defined sustainability goals for the data center by optimizing the joint operation of these resources. The challenges lie in the physics limitations of SMR operation and the prediction errors of renewable energy from the main grid.

1.3 Research Framework

This thesis addresses the critical challenge of carbon-aware and robust energy systems scheduling under uncertainty, a fundamental problem in supporting the global energy transition. Decarbonization efforts span diverse energy use domains. Accordingly, this research investigates this core scheduling problem through the lens of three distinct, yet interconnected, system focus areas: Multi-Energy Building Integrated Energy Systems (BIES), Thermostatically Controlled Loads (TCLs), and Data Center Energy Systems (DCES). While each system presents unique operational characteristics and specific low-carbon objectives (e.g., net-zero for BIES, load flexibility for TCLs, 24/7 carbon-free energy for DCES), they share fundamental complexities. These include: accurately accounting for and integrating carbon considerations into scheduling decisions; managing inherent uncertainties in energy forecasts (like loads and renewable generation) and carbon intensity predictions; and developing computationally effective optimization strategies that provide robust operational guarantees against these uncertainties. The integrated nature of this research lies in its unified methodological foundation: applying and advancing optimization-based scheduling techniques that explicitly incorporate carbon awareness and, crucially, focusing on advanced techniques for handling uncertainty. A central contribution across all investigations is the development and application of sophisticated methods to ensure scheduled operations are robust and reliable in the presence of prediction errors. This general framework is tailored and applied to the specific challenges of each system as follows:

1. For multi-energy BIES, addressing the goal of net-zero operations under significant forecast errors (loads, renewable generation, carbon intensity), we develop a learning-based robust optimization approach. This method is specifically designed to handle these diverse uncertainties while guaranteeing statistical feasibility of the scheduling decisions.

2. For TCLs, supporting carbon-aware scheduling in the face of uncertainties like building base loads and outdoor temperatures, we propose a novel bilevel optimization reformulation strategy. This technique directly addresses common limitations of conventional approaches to handling uncertainty, such as over-conservatism and computational intractability, particularly in the context of

chance constraints.

3. For DCES, targeting the ambitious goal of 24/7 CFE reliability despite uncertainty in renewable energy generation forecasts, we introduce an algorithm leveraging multi-variable conformal prediction. This method utilizes covariate information to construct statistically valid and adaptive uncertainty sets for renewable forecasts, thereby enhancing the robustness of the scheduling.

4. For the nuclear-powered DCES, 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.

This structured approach allows for the exploration of common challenges and methodological advancements in carbon-aware, robust scheduling under uncertainty across diverse energy domains, demonstrating the broad applicability of the proposed framework and techniques.

1.4 Contributions

This thesis makes significant contributions to the field of carbon-aware and robust energy system scheduling under uncertainty, addressing critical challenges across diverse domains to support the global energy transition. Specifically, the primary contributions of this work are presented through the following:

1. We develop a novel carbon-conscious and robust learning-based optimization framework for the efficient operation of BIES components. This framework explicitly addresses multiple forecast uncertainties (loads, renewables, carbon intensity) by constructing appropriate uncertainty sets and provides crucial statistical feasibility guarantees, ensuring reliable net-zero operations validated

through extensive numerical studies.

2. We introduce a novel bilevel optimization reformulation strategy for the data-driven distributionally robust chance-constrained (DRCC) scheduling of TCLs under outdoor temperature and base load uncertainties. This approach effectively mitigates issues of over-conservatism and computational intractability common in existing methods, demonstrating substantial improvements in reducing regret, energy costs, and carbon emissions in simulation studies using real-world data.

3. We propose a robust optimization framework leveraging multi-variable conformal prediction for carbon-conscious workload scheduling in data centers, aiming for 24/7 carbon-free energy (CFE). This method provides statistical feasibility guarantees by integrating covariate information to construct statistically valid and adaptive uncertainty sets for renewable generation forecasts, enabling more reliable operations, improved violation rate control, and significant cost reductions compared to traditional methods.

4. We formulate a day-ahead data center workload planning with LF-SMR 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 predefined requirement on the combined share of renewable and nuclear carbon-free energy with statistical confidence.

5. The performance of all three proposed techniques was evaluated through extensive experiments. These experiments involved comparison against representative benchmarks, and the results demonstrate the effectiveness of the proposed

algorithms.

1.5 Thesis Organization

The rest of this thesis is organized as follows.

1. **Chapter 2 (Background and Literature Review)**: This chapter provides the necessary foundational knowledge for the subsequent research. It includes a review of relevant background information on advanced low-carbon technologies in energy systems and a comprehensive survey of the literature concerning frameworks and algorithms for handling uncertainty in optimization-based energy system scheduling.

2. **Chapter 3 (Robust Net-Zero Scheduling of Multi-energy Building Integrated Energy Systems):** This chapter presents the detailed formulation and development of the proposed learning-based robust optimization framework for achieving net-zero operation in multi-energy building integrated energy systems under uncertainty.

3. **Chapter 4 (Carbon-aware Scheduling of Thermostatically Controlled Loads):** This chapter focuses on the carbon-aware scheduling of TCLs. It details the formulation of the data-driven DRCC problem under uncertainties arising from outdoor temperature and base load forecast errors.

4. **Chapter 5 (24/7 Carbon-Free Data Centers Operation Scheduling):** This chapter investigates the problem of achieving 24/7 CFE operation for DCES through carbon-conscious workload scheduling under the prediction errors of renewable energy generation.

5. **Chapter 6 (Data Centers Workload Planning with Small Modular Reactors):** This chapter investigates the problem of jointly coordinating data-

center workload planning and load-following small modular reactors operations under the main grid renewable energy generation prediction errors.

6. **Chapter 7 (Conclusions and Suggestions for Future Research):** This chapter concludes the thesis by summarizing the key findings and contributions of the research works presented. It also discusses promising directions for future research in carbon-aware and robust energy system scheduling under uncertainty.

Chapter 2

Background and Literature Review

2.1 Background of Low-carbon technologies

2.1.1 Carbon Capture, Utilization, and Storage

Carbon capture, utilization, and storage (CCUS) technologies represent crucial strategies for mitigating carbon emissions. These approaches involve the capture of carbon dioxide, typically from large point sources, which is subsequently directed down one of two primary pathways: storage or utilization. Carbon capture and storage (CCS) involves compressing and transporting the captured CO_2 for permanent sequestration by injecting it into suitable deep underground geological formations [31]. This pathway has demonstrated effectiveness in preventing CO_2 release into the atmosphere. Carbon capture and utilization (CCU), alternatively, refers to applications where the captured CO_2 is used as a feedstock. CCU can involve using CO_2 either directly (without chemical transformation) or indirectly (by converting it) into various commercially valuable products [26]. Established direct utilization pathways include its use in the fertiliser industry for urea manufacturing and for enhanced oil recovery. Furthermore, new utilization pathways focused on the production of CO_2-based synthetic fuels, chemicals, and building aggregates are currently under development and gaining significant interest.

While traditional CCS efforts have primarily targeted large point sources such as industrial facilities and power plants, distributed CCS solutions for buildings are increasingly recognized as a viable pathway for decarbonizing urban infrastructure. Within the building environment, post-combustion CCS, which involves separating CO_2 from flue gases produced by natural gas combustion, is considered the most feasible technique due to its adaptability. These post-combustion units can often be retrofitted into existing natural gas heating systems and furnaces with minimal modifications to the building's infrastructure.

Academic literature has explored the integration of carbon capture technologies within buildings. Studies have discussed the market feasibility of installing CCS at the building level [131], detailed the design of building-integrated carbon capture (BICC) for retrofitting existing mechanical systems and highlighted its potential role in offsetting emissions [147], modeled the deployment of CCS alongside energy efficiency and demand response measures for zero-emission targets [157], and conducted comparative life cycle assessments evaluating the environmental and economic performance of building-level CCS, demonstrating its potential in regions reliant on fossil fuels [114].

Practical applications of building-level CCS are also emerging. For instance, FortisBC Energy initiated a pilot program in Canada in 2017, deploying and testing CCS for natural gas-fueled building heating systems in various commercial buildings [2, 18]. Their latest installation at Southridge School in 2023 demonstrated the potential for substantial CO_2 reductions (up to 5,400 kg per unit per year) alongside reported energy consumption decreases (up to 10%) [8]. Concurrently, commercial carbon capture products and services are becoming readily available. CarbonQuest, for example, offers modular systems designed for capturing CO_2 from boilers, combined heat and power (CHP) systems, and fuel cells across diverse building types, including offices, campuses, and multifamily residences. Their scalable systems range from capturing 500 to 16,800 tons of CO_2 annually for typical commercial buildings, with larger configurations exceeding 100,000 tons per year in campus settings [20, 19]. Deployments include the first on-site system in a Manhattan multifamily building in 2021 [22, 13]. Furthermore, innovative approaches like HVAC-integrated direct air capture (DAC) are being explored, such as Soletair Power's technology, which captures ambient CO_2 using sorbents, with units capable of capturing approximately 20 tons/year and offering advantages like low noise and renewable energy compatibility, as demonstrated in pilot projects in Finland and Dubai [30].

2.1.2 Grid Carbon Intensity

Electrical grids typically receive power from a diverse portfolio of energy sources, including conventional fossil fuels (e.g., coal, natural gas), nuclear power, and renewable resources (e.g., wind, solar, hydro), among others. These sources exhibit significant variations in their associated greenhouse gas emissions profiles; burning fossil fuels releases substantial amounts of CO_2, while renewable sources generally have much lower operational carbon footprints.

The grid carbon intensity (GCI) quantifies the greenhouse gas emissions, typically measured in terms of CO_2 equivalents, associated with the generation and delivery of electricity. It is commonly expressed in units of grams of CO_2 per kilowatt-hour (gCO_2/kWh) and reflects the environmental impact of electricity consumption based on the grid's prevailing energy mix [25]. A key characteristic of GCI is its dynamic nature; it varies significantly both temporally (e.g., hourly fluctuations driven by changes in generation dispatch and demand, as well as seasonal shifts) and spatially (reflecting the distinct energy generation mixes of different regional grids).

It is also important to distinguish between different types of GCI data. Average GCI represents the weighted average emissions intensity across all energy sources currently supplying the grid. In contrast, marginal GCI refers to the emissions intensity of the

Table 2.1: Carbon-emission rate (g/kWh) for different sources

Emission factors	Coal	Oil	Natural gas	Nuclear	Solar	Wind	Hydro	Other	Biomass	Geothermal
Lifecycle	820	650	490	12	45	11	24	700	230	38
Direct	760	406	370	0	0	0	0	575	0	0

specific generating unit that responds to changes in demand (often the most flexible, which historically tends to be fossil-fuel based). The choice between average and marginal GCI depends on the specific application and the nature of the scheduling or operational decision being made.

GCI is calculated based on the carbon emission factor (CEF) of different energy sources. Table 2.1 shows the CEF of different energy sources [120]. CEF represents the amount of carbon dioxide equivalent emitted into the atmosphere per unit of electricity generated. Generally, non-renewable sources, such as coal and natural gas, exhibit substantially higher CEFs compared to renewable sources like wind and solar power. The CEF for a given source can be calculated based on different accounting scopes, primarily distinguishing between operational and lifecycle emissions:

- Direct Emission Factors: These factors account for the operational emissions released during the conversion of a source into electricity. They are primarily relevant for scope 2 emissions accounting [3].

- Lifecycle Emission Factors: These factors provide a more comprehensive view by including operational emissions alongside upstream and downstream emissions associated with the source's supply chain, infrastructure development, manufacturing, and decommissioning. They are relevant for scope 3 emissions accounting [4].

Furthermore, the specific CEF value for a given energy source can vary depending on factors such as the specific power plant technology, operational efficiency, and the quality of the fuel used (e.g., different types of coal). It is important to note that accurately determining accurate and site-specific CEFs can be a complex analytical challenge.

Understanding and tracking GCI is essential for multiple aspects of decarbonization [27]. Firstly, it provides a clear delineation of the environmental impact associated with electricity consumption at any given moment. Secondly, monitoring temporal changes in GCI serves as a direct measure of progress in decarbonizing the electricity sector itself. Furthermore, this metric is crucial for informing policy decisions and guiding investments towards cleaner energy technologies and grid modernization. Finally, insights into GCI can empower consumers and energy managers to make more sustainable energy choices, such as optimizing electricity consumption to align with periods when the grid is supplied by lower-carbon sources.

2.2 Carbon-aware Operation Scheduling

2.2.1 Low-carbon Operation in BIES

Recent research has significantly advanced the optimization of BIES with explicit consideration of carbon emissions, often through the integration of various low-carbon technologies. Notably, studies have explored the integration of technologies such as CCS and Power-to-Gas (P2G) into energy system optimization. Li et al. [109], for example, demonstrated how integrating CCS and P2G can enhance the stability of conventional energy systems while addressing energy demand and renewable generation uncertainties. Xiao et al. [178] investigated a multi-energy combined heat and power (CHP)-based microgrid incorporating a P2G facility for CO_2 reduction, considering the potential for captured CO_2 utilization and sale. Furthermore, a hydrogen-based framework combining P2G and CCS was proposed by Wu et al. [176], highlighting carbon trading mechanisms as a means to improve the cost-effectiveness of CCS by converting emission reductions into financial incentives, a practice already being implemented in some regions within the building sector [175]. These contributions collectively mark significant progress in integrating advanced technologies and tracking carbon flows within building energy management.

While these studies represent important advancements in optimizing BIES for decarbonization, a critical limitation persists in these works: the common reliance on fixed, average emission factors for grid electricity. These approaches inherently overlooks the substantial temporal and spatial variability in the grid's carbon intensity, which is dynamically influenced by the changing mix of generation sources (e.g., higher reliance on renewables at certain times/locations). As the building sector undergoes rapid decarbonization, such simplified carbon accounting can lead to suboptimal operational strategies that fail to leverage opportunities to consume electricity when its carbon intensity is low or shift load away from high-intensity periods. Therefore, there is a compelling need for more accurate carbon emission modeling and accounting within BIES operational strategies that explicitly consider and exploit the spatial and temporal dynamics of the grid's carbon intensity.

2.2.2 Carbon-aware Demand Response

Carbon-aware demand response (CADR) is an advanced demand-side management strategy that optimizes energy consumption based on real-time grid carbon intensity (gCO2/kWh) and electricity prices. Unlike traditional demand response, which focuses solely on cost or grid stability, CADR integrates environmental metrics to reduce carbon emissions while maintaining efficiency. Research in this area explores various applications and challenges, from optimizing individual user behaviour and integrating flexible loads into grid operations to applying CADR principles to specific, energy-intensive sectors. Several studies highlight the importance and potential of incorporating carbon awareness into demand response mechanisms across different scales and load types. Chen *et al.* delves into the carbon-aware demand response

(C-DR) paradigm from the perspective of individual users in [59], aiming to minimize their carbon footprints by optimally scheduling flexible loads like deferrable devices and thermostatically controlled loads. This work integrates the user-level C-DR into a grid-level optimal power dispatch model, demonstrating how user-driven carbon consciousness can contribute to low-carbon power system operation. Addressing the significant energy demands of specific sectors, Xing *et al.* investigate implementing CADR in data centers by adjusting their power usage based on grid signals in [179]. The paper proposes the Carbon Responder framework to reduce datacenters' carbon footprint by adjusting power usage based on grid carbon signals. A key focus is managing diverse workloads and ensuring fair power allocation while respecting performance requirements, showcasing substantial carbon reduction benefits compared to traditional methods. Large, flexible loads like Electric Vehicles (EVs) also present a prime opportunity for CADR. Wang *et al.* addresses the challenge of increased carbon emissions from growing EV charging demand in [172]. The paper develops a carbon-aware methodology to quantify and manage aggregate EV power flexibility in real-time under uncertainties in both EV behaviour and grid carbon intensity. This approach effectively controls emissions and highlights the role of intelligent EV charging in grid decarbonization.

These papers illustrate the breadth of CADR research, covering foundational concepts for individual users and grid integration, as well as practical applications and benefits within energy-intensive sectors like data centers and major flexible loads like EVs. They underscore the potential of shifting energy consumption based on grid carbon intensity to achieve significant environmental benefits across diverse domains.

2.2.3 24/7 CEF Operations for Data Centers

The pursuit of sustainability in data centers is increasingly guided by ambitious targets, notably net-zero emissions and 24/7 carbon-free energy (CFE). While both aim to reduce environmental impact, they represent distinct approaches. Net-zero emissions focus on balancing greenhouse gas emissions annually across all scopes, permitting the use of offsets and carbon removal technologies to negate residual emissions [28]. In contrast, 24/7 CFE is a more rigorous target specifically for electricity consumption (primarily Scope 2), demanding that every kilowatt-hour consumed is matched by carbon-free generation on the local grid, during the same hour it is consumed [17]. As 24/7 CFE requires direct procurement and real-time management of clean energy, moving beyond annual energy certificate purchases or global balancing methods characteristic of net-zero.

Existing research has explored various aspects of green data centers, including optimizing economic performance with onsite renewable energy [77], strategically allocating workloads based on green energy availability for profit [98], and implementing advanced energy management and demand response programs to reduce carbon emissions, often involving co-optimization of server operations and leveraging integrated renewables and storage [64, 100]. While these contributions provide valuable insights into renewable integration and demand-side flexibility, they typically align with broader, less temporally stringent goals.

Achieving 24/7 CFE, however, presents unique and significant operational challenges for data centers. These include overcoming the intermittency of renewable generation and the current limitations of energy storage technologies needed to cover continuous demand, navigating geographical constraints related to the availability of local carbon-free resources, and managing the increasingly high and dynamic power requirements of modern workloads like AI [5]. Although the concept of 24/7 carbon-free data center operations has been introduced in [33], which discusses long-term investment strategies, there is a critical lack of research addressing the operational-level implementation and real-time control mechanisms necessary to dynamically match data center electricity demand with available carbon-free energy sources throughout all hours of the day. Addressing this gap in real-time operational management is essential for realizing true 24/7 CFE data center operations.

2.2.4 Nuclear Energy as a Solution for Data centers

Nuclear energy presents a compelling option for powering data centers, offering a reliable and low-carbon energy source. Traditionally, Nuclear Power Plants (NPPs) operate in a continuous baseload mode, consistently producing their maximum rated capacity. However, NPPs are technically capable of more flexible operation, known as load-following. In this mode, plants can adjust their power output to match varying demand, contributing valuable services like frequency regulation and operating reserves to enhance grid reliability. Several countries, including France, Germany, the United States, Belgium, and Switzerland, possess significant expertise in flexibly operating their nuclear fleets [93].

When considering nuclear solutions for data centers, options include large utility-scale plants, Small Modular Reactors (SMRs), and microreactors [151]. A typical large NPP, generating around 1 gigawatt, generally exceeds the power requirements of a single data center. Consequently, SMRs, with capacities up to 300 megawatts, are often considered more suitable. SMRs are characterized by their smaller size and inherent flexibility [116]. This reduced footprint facilitates greater deployment flexibility, and their modular nature simplifies construction and installation processes.

However, achieving flexible operation, whether with SMRs or traditional NPPs, is subject to technical limitations rooted in reactor physics. A primary constraint is the issue of xenon poisoning: the buildup of Xenon-135, a strong neutron absorber, which occurs after a reduction in power and inserts negative reactivity into the core, complicating subsequent power increases or stable low-power operation.

2.3 Optimization under Uncertainty

2.3.1 Robust Optimization

Robust optimization (RO) has been a prominent approach for managing uncertainties in complex systems such as building integrated energy systems (BIES) and energy hub operations. The core principle of RO is to find solutions that remain feasible and optimal for all possible realizations of uncertain parameters within a defined uncertainty set. This method is particularly useful when precise probability distributions of uncertainties are unknown or difficult to obtain. Examples from prior research include the development of a two-stage robust scheduling method for multi-energy systems to enhance flexibility against wind energy variability [181]. In the context of TCL scheduling, RO methods have also been examined for handling forecasting errors of ambient temperature [60]. However, a significant drawback of traditional RO is its potential to yield overly conservative outcomes, which can diminish the overall quality and economic efficiency of the solution. While data-driven set-based RO methods have been proposed to mitigate this over-conservativeness [108], they, along with other data-driven approaches, have been noted to lack statistical feasibility guarantees [89].

2.3.2 Distributional Robust Optimization

Distributionally robust optimization (DRO) emerges as an advanced method designed to handle uncertainties without requiring precise knowledge of the underlying probability distributions. Instead of assuming a single distribution, DRO optimizes over a family of possible distributions, known as an ambiguity set, which is constructed based on available empirical data. This approach offers a middle ground between stochastic optimization (which requires exact distributions) and traditional robust optimization (which can be overly conservative). DRO has been applied to address specific uncertainties, such as PV output uncertainty using a method based on Wasserstein distance [91], and in DRO-based integrated energy system optimization [176]. Furthermore, the concept is extended to distributionally robust chance-constrained (DRCC) methods, which facilitate decision-making over an ambiguity set, particularly useful in contexts like tackling uncertainties from heating demand, outdoor temperature [58], and ambient temperature forecasting errors in TCL scheduling [183]. While DRO and its variants like DRCC address the limitation of requiring precise distributions, data-driven DRO methods, similar to data-driven RO and chance constraints, still face challenges regarding statistical feasibility guarantees [89].

2.3.3 Chance-constraints Programming

Chance-constrained programming (CCP) is a method employed to ensure that a set of constraints is satisfied with a certain minimum probability. This approach acknowledges the inherent randomness of uncertain parameters and allows for a controlled level of constraint violation, striking a balance between robustness and conservativeness compared to stochastic and robust optimization methods. CCP has been utilized for handling uncertainties in loads and renewable sources like wind and PV, often through scenario-based joint chance constraint approaches [178]. In the context of TCL scheduling, CCP is used to guarantee a certain probability of occupant comfort, even under challenging conditions [164]. A key limitation of standard CCP, however, is its requirement for explicit knowledge of the probability distributions for the random variables involved [54]. To overcome this, distributionally robust chance-constrained (DRCC) methods have been introduced [58, 183], leveraging ambiguity sets derived from data to make decisions robust over a family of distributions. Solving complex CCPs and DRCCs presents computational challenges due to their non-differentiability and non-convexity. Data-driven DRCC control problems, particularly in TCL applications, have explored various solution approaches, including the Conditional Value at Risk (CVaR) approximation [71, 173]. CVaR offers advantages in terms of linearity and incorporating supporting information but can lead to increased conservativeness due to its focus on the severity of violations [126, 44]. An alternative is exact deterministic reformulations for data-driven CCPs within Wasserstein balls [61], but these often transform the problem into computationally prohibitive Mixed-Integer Programs (MIPs) for larger scales [62]. While data-driven chance-constraints methods have sought to address over-conservativeness [192], they, alongside data-driven RO and DRO, are noted for lacking statistical feasibility guarantees [89].

2.3.4 Contextual Robust Optimization

A fundamental limitation of traditional robust optimization and many data-driven uncertainty handling methods, including certain stochastic optimization approaches, is their tendency to overlook the inherent relationship between observable contextual information (covariates) and the specific characteristics of prediction uncertainty. This oversight often results in the use of static, worst-case uncertainty sets or fixed probability distributions that do not adapt to real-time conditions. Consequently, decisions derived using these methods can be overly conservative, as they fail to leverage information that could provide a more accurate, context-dependent view of the uncertainty.

In reality, the structure and magnitude of uncertainty are often not uniform but vary significantly depending on the context provided by these covariates. For example, the distribution of prediction errors for a load forecast may differ considerably based on factors like the time of day, weather conditions, or even the value being predicted itself. Therefore, effectively addressing this limitation requires methods capable of explicitly leveraging covariate information to construct statistically rigorous and adaptive uncertainty sets.

The field of data-driven contextual optimization has emerged to tackle decision-making problems under uncertainty by explicitly leveraging such available contextual information [145]. Within this domain, contextual robust optimization (CRO) presents a compelling approach, particularly suited for risk-averse decision-makers. Unlike approaches that require modeling precise conditional probability distributions (as in contextual stochastic programming [145], which may focus on risk-neutral objectives or necessitate distribution knowledge for chance constraints), CRO directly utilizes covariates to construct conditional uncertainty sets.

A core component of CRO is the application of machine learning techniques to analyze covariates and extract underlying patterns. Methods such as the k-nearest neighbors (kNN) model [106], conformal prediction [130], and neural networks [51] are employed for this purpose. This analysis enables the construction of an uncertainty set that is conditional on the specific observation of covariates at the time of decision-making. This conditional uncertainty set is designed to encapsulate the actual realization of uncertainty with a high probability given the observed context. Crucially, because the set is tailored to the specific context, it is typically smaller and less conservative than a generic uncertainty set derived without considering covariates, thereby improving solution quality while maintaining robustness guarantees.

Chapter 3

Robust Net-Zero Scheduling of Multi-energy Building Energy Systems

3.1 Introduction

Buildings contribute significantly to carbon neutralization due to their significant energy consumption, carbon emissions, and their potential to lead in the adoption of energy-efficient technologies and renewable energy sources. For instance, nearly a third of U.S. greenhouse gas emissions are attributable to 130 million homes and commercial buildings in America, which use 40% of the nation's energy and 75% of its electricity [167]. To achieve the urgent climate goal, the existing design of building integrated energy system (BIES) needs to evolve in a climate-responsible manner to increasingly aggregate zero/low-emission energy technologies and optimize energy efficiency, while meeting the demand of heat, power, and cooling [72, 190]. Existing research has shown that a climate-smart BIES not only serves as the energy backbone of buildings but also acts as an innovative platform for the integration of various new technologies and strategies within the building environment to cut carbon emissions. For instance, empirical findings in [166] indicated that the implementation of technologies such as photovoltaic (PV) and energy storage systems (ESSs) in BIES could potentially decrease CO_2 emissions by 20%.

Despite the potential benefits offered by BIES, its optimal scheduling is not straightforward. The first research challenge we are facing is: *How to perform synergic carbon-aware coordination of* ***on-site*** *components and technologies while optimizing* ***off-site*** *electricity usage from the utility grid, to achieve net-zero emission?* BIES comprises multiple on-site components, like power-to-gas (P2G) and carbon capture systems (CCS), each presenting a method to harness renewable energy and reduce carbon footprints. The coordination between these diverse technologies and the smart use of off-site, low-carbon electricity is essential. As the share of renewable energy in the grid increases, the carbon intensity of electricity fluctuates, making it crucial to schedule energy-intensive operations during times of lower carbon intensity to minimize the carbon emission of buildings.

Another major challenge is: *How do we handle uncertainties inherited in BIES operations?* While renewable sources such as wind and solar generation do not create any carbon emissions, their intermittent and stochastic nature poses severe challenges to BIES operation and safety. Energy prediction has thus become a key element to accurately forecast future operating conditions using historical operation data and machine learning algorithms [143]. However, all prediction models would inevitably inherit errors. The sources of uncertainty for optimal BIES design stem from load prediction, such as cooling/heating and electrical load, as well as renewable energy production [149]. The presence of the uncertainty factor may adversely affect the performance of the model [191] and challenge the practical application of deterministic methods [170]. These uncertainties should be systematically considered during the energy management decision-making process involved in BIES to ensure its feasibility under various operating scenarios [196].

Motivated by the aforementioned research gaps, in this paper, we develop a novel carbon-aware and learning-based robust operation framework for a BIES comprising of a range of technologies, such as CCS, P2G, energy storage system, combined heat and power, absorption chiller, and electric chiller. To take carbon emissions into account accurately, we incorporate carbon trading and time-varying carbon intensity of the regional grid, along with zero-emission technologies, into the formulation of the BIES scheduling problem formulation to enable well-informed decision-making. Then, we incorporate the uncertainties related to electricity, heating, cooling load forecast errors, PV power generation, and carbon intensity into the scheduling problem through chance-constrained optimization, which can be converted into a corresponding robust optimization problem. We propose a two-phase approach comprising shape learning and shape calibration in order to acquire an optimal uncertainty set that ensures favorable results with a statistical feasibility guarantee. The proposed approach was tested based on synthetic and real-world data. Simulation results indicate that the proposed learning-based robust optimization approach yields an average 8.2% reduction in cost and net-zero emission building, compared with performance benchmarks, while ensuring stable feasibility guarantees.

The contributions of this paper are summarized as follows:

1. We develop a carbon-aware formulation for optimal BIES operation incorporating low-carbon technologies, carbon accounting, and carbon trading. Our work pioneers the integration of precise carbon emissions modeling into the BIES formulation to facilitate the zero-carbon transition of the building sector, distinguishing it from literature that concentrates primarily on minimizing operational costs.

2. We develop a novel two-stage learning-based uncertainty set construction algorithm that is capable of ensuring statistical feasibility. Through theoretical analysis and experimental verification, we demonstrate that the proposed algo-

rithm guarantees that the solutions consistently adhere to satisfy the building energy and carbon consumption balance at a given confidence level.

3. We conduct extensive simulation studies using both synthetic and real-world datasets, showcasing the performance of the proposed approach, compared with existing performance benchmarks. The findings indicate that our approach demonstrates improved control over the violation rate and yields lower costs.

Table 3.1 provides a detailed comparison between this work and existing research efforts. It is worth noting that uncertainties persist without consideration of carbon-related factors in the building environment, and the proposed two-phase approach can work with other forms of uncertainties. However, we are incorporating carbon accounting and modeling in this manuscript to underscore the urgency and significance of transitioning towards more sustainable and environmentally friendly building practices. Through the "carbon-aware" lens, we aim to provide insights that can inform and drive effective strategies for achieving carbon reductions in the built environment.

The remainder of this paper is organized as follows. Section 3.2 discusses the related work. Section 5.3 presents the model of the carbon capture system as well as the model of BIES. Section 3.4 illustrates the solution methodology to deal with the uncertainty. The numerical studies are carried out in Section 4.6, and finally, the conclusions are drawn in Section 5.7.

3.2 Related Work

3.2.1 Carbon Capture in BIES

CCS has proven to be an effective solution for reducing CO_2 emissions by capturing carbon directly at the source. The captured carbon can then be either stored long-term in suitable sites or converted into commercially useful by-products [107].

In the building environment, post-combustion CCS, which refers to the technology that separates CO_2 from the flue gases produced after the combustion of natural gas, is considered the most feasible technique. Post-combustion CCS units can be integrated into existing natural gas combustion systems and furnaces through retrofitting, requiring minimal modifications to building infrastructure. To date, there is an extensive body of literature on the integration of carbon capture technologies within building environments. For instance, the feasibility of installing CCS at the building level has been discussed in [131] from the market assessment perspective. The design of building integrated carbon capture (BICC) for retrofitting existing mechanical systems within buildings was discussed in detail in [147], highlighting the significant role BICC can play in offsetting carbon emissions from different types of buildings. The modeling of CCS to deploy along with energy efficiency measures and demand response to achieve zero-emission was discussed in [157]. A comparative performance assessment was conducted in [114] to evaluate the environmental and economic performance of building-level CCS over their life cycle, showcasing the great potential for building-level CCT in regions that depend on fossil fuels to generate electricity.

In practical applications, CCS for natural gas-fueled building heating systems was first deployed and tested by FortisBC Energy in a pilot program in Canada in 2017 [2]. The participants of this program included various commercial buildings, including LUSH Cosmetics Headquarters, Baptist Housing, Starlight Housing, Richmond Centre, and Richmond School District [18]. Their latest small-scale CCS installation project was at Southridge School in 2023 [8]. It was reported that the carbon capture process could help commercial customers reduce up to 5,400 kilograms of CO_2 reduction per unit per year while decreasing energy consumption by up to 10% [8]. Furthermore, commercial carbon capture products and services, such as the solutions provided by CarbonQuest [20], have been made readily available for various building types. For example, it is reported that their larger-size Distributed Carbon Capture products can be bundled together for CHP solutions to capture over 100,000 tons of CO_2 in campus settings, including universities, hospitals, and multi-building complexes [19]. Meanwhile, their Building Carbon Capture products can be deployed in commercial buildings, such as offices and mixed-use buildings, to capture 500 to 16,800 tons of CO_2 per year [19]. So far, their products have been reportedly deployed in 5 locations, with its first on-site carbon capture system installed in a multifamily building located in Manhattan's Lincoln Square neighborhood in 2021 [22][13].

At the same time, countries around the world are increasingly recognizing the significance of CCT in mitigating climate change. For instance, in the United States, the 45Q Tax Credit included in the Inflation Reduction Act (IRA) provides financial incentives and facilitates the acceleration of the demonstration and deployment of CCT in commercial applications [14]. Similarly, the European Union's Innovation Fund supports the demonstration of innovative low-carbon technologies including CCS [9]. In addition, the United Kingdom, Australia, and China have also implemented specific policies and subsidies to promote CCS [16]. These measures are crucial for expediting the commercialization of CCS technologies, which are essential for attaining global climate targets.

3.2.2 Low-carbon Operation of BIES

Recent studies have significantly advanced the optimization of BIES with carbon emissions taken into account by utilizing a combination of different carbon-free technologies. In the research literature, Li et al. [109] demonstrate the integration of CCS and P2G technologies into conventional energy systems, enhancing stability and addressing energy demand and renewable generation uncertainties. Xiao et al. [178] explore a multi-energy CHP-based microgrid for smart cities, incorporating a P2G facility to reduce CO_2 emissions and considering the sale and utilization of captured CO_2. A hydrogen-based framework that combines a refined P2G system and CCS is developed in [176], offering a novel approach to energy system design. It also high-

Table 3.1: Comparison between this study and existing approaches

Ref.	Carbon-Aware Aware	Grid Emission Factor	Uncertainty Modeling	Distribution Assumption-Free	Violation Rate Guarantee	Statistical Feasibility Guarantee
[148]	×	×	Stochastic	×	×	×
[192]	×	×	Chance constraint	×	×	×
[108]	√	×	Robust optimization	√	√	×
[109]	√	×	Stochastic	×	×	×
[178]	√	Constant	Chance constraint	×	×	×
[176]	√	Constant	DRO	√	√	×
Our work	√	Real-time	Learning-based robust	√	√	√

lights that to tackle the concern regarding the potential lack of cost-effectiveness in CCS applications, building owners can additionally engage in carbon trading mechanisms. These mechanisms convert emission reductions into financial gains. Some countries and regions have already implemented or are in the process of integrating carbon trading into the building sector [175]. *Research Gaps:* While these contributions mark significant progress in tracking carbon flows within the building, they often rely on fixed emission factors for grid electricity, overlooking the variability in carbon intensity due to diverse energy sources. It is evident that this status quo can no longer meet the demands of the building sector, which is undergoing rapid and comprehensive decarbonization. There is a compelling need for more accurate carbon emission modeling and accounting to determine the operational strategy of BIES, taking into account the spatial and temporal dynamics of the grid's carbon intensity.

3.2.3 Handling Uncertainties in Multi-energy Systems

Various methods have been investigated in prior research to handle uncertainties in BIES and energy hub operations. Representative methods include chance-constrained optimization, robust optimization, stochastic optimization, and lately, distributionally robust optimization. For instance, Yan et al. [181] develop a two-stage robust scheduling method for multi-energy systems, enhancing flexibility against wind energy variability. To account for the uncertainty from PV output, a distributional robust optimization (DRO) method based on Wasserstein distance is developed in [91]. A distributional robust chance-constrained (DRCC) method has been introduced in [58] to tackle uncertainties from heating demand and outdoor temperature. Xiao et al. [178] utilizes a scenario-based joint chance constraint approach for handling uncertainties in loads and renewable sources like wind and PV. These studies primarily employ stochastic or robust optimization methods. Stochastic optimization requires precise probability distributions of uncertainties, while robust optimization might lead to overly conservative outcomes, potentially diminishing solution quality. Although the data-driven set based RO method presented by [108], the chance-constraints method proposed by [192], and DRO-based IES optimization proposed in [176] address the issue of over-conservativeness, they lack statistical feasibility guarantees [89]. *Research gaps:* Existing methods primarily employ samples to estimate the characteristics of random variables, such as their distributions and uncertainty sets. However, there are often discrepancies between the estimated distributions and the true distributions of these random variables, leading to inaccuracies in predictions, which can result in the violation of constraints and suboptimal solutions. To address this critical challenge, the concept of statistical feasibility emerges as an effective solution to ensure that the solutions obtained from an optimization problem with chance constraints are feasible across different samples with a certain level of confidence. Statistical feasibility allows decision-makers to derive a solution that is robust from a statistical perspective, thereby eliminating the need for assumptions about sample distribution or resampling techniques often found in the literature. To guarantee statistical feasibility, the uncertainty sets need to be properly calibrated from data, so that we can quantify the errors to be involved in the optimization process in a probabilistic way.

3.3 Model Formulation

This section outlines the carbon-related and other components of the BIES, crucial for achieving carbon neutrality. Components include Combined Heat and Power (CHP), Energy Storage System (ESS), Absorption Chiller (AC), Gas Boiler (GB), Electric Chiller (EC), and purchased electricity from the grid. Figure 4.1 illustrates the system structure and energy flow. The BIES aims to minimize costs, including energy purchases and carbon credits, through optimal operation, balancing cost-effectiveness with low carbon emissions.

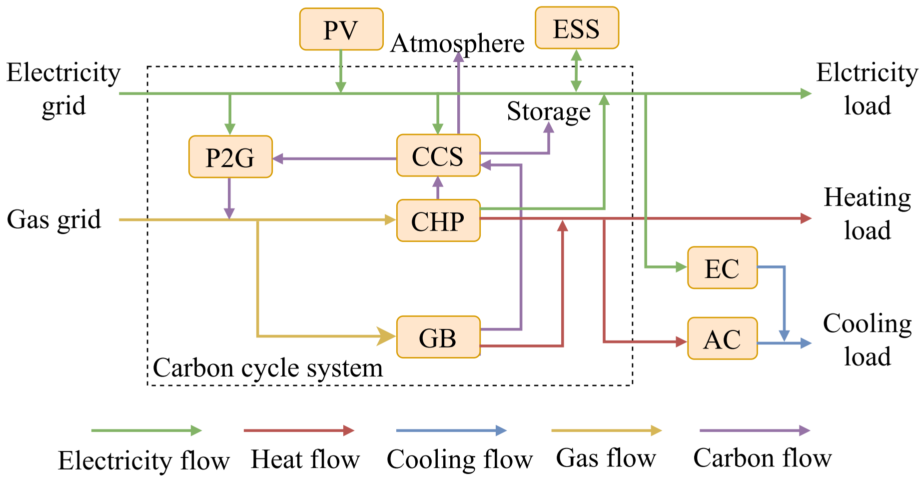


Figure 3.1: The brief structure of BIES and energy flow

3.3.1 Model of Carbon Cycle System (CCS)

To mitigate the carbon emissions of the system, it is essential to quantify these emissions accurately. This subsection introduces the carbon cycle system, which includes the processes of CO_2 production, capture, and utilization. Following this, we present a model for carbon accounting that provides a comprehensive framework for understanding and reducing the carbon footprint of the BIES.

Carbon production: In BIES, the CO_2 produced consists of two parts: the direct emission and the indirect emission. The direct carbon emission comes from the consumption of natural gas, and we have

$$E_{CHP,t} = \delta_g F_{CHP,t}, \tag{3.1}$$

$$E_{GB,t} = \delta_g F_{GB,t}, \tag{3.2}$$

where $E_{CHP,t}$ and $E_{GB,t}$ are the amount of carbon emission from CHP and GB, respectively. δ_g represents the CO_2 emission factor of natural gas. $F_{CHP,t}$ and $F_{GB,t}$ denote gas consumption of CHP and GB.

The indirect carbon emission comes from the power purchased from the main grid, and we have

$$E_{virtual} = \sum\nolimits_{t=1}^{T} CI_t * P_{buy,t}, \tag{3.3}$$

where $P_{buy,t}$ is the purchased power from the main grid. CI_t is the carbon intensity of the main grid at time t. Calculating CI_t involves analyzing the mix of energy sources in the grid and their respective emission factors. Specifically, CI_t is determined as follows [193]:

$$CI_t = \frac{\sum_{k \in \mathcal{E}} ef^k \times M_t^k}{\sum_{k \in \mathcal{E}} M_t^k}, \tag{3.4}$$

where $\mathcal{E}$ as the set of energy sources. M_t^k represents the amount of electricity produced by source k at time t. The carbon emissions associated with electricity generation vary across different energy sources. Each source k has an associated carbon emission factor, ef^k, indicating the emissions per unit of electricity generated by this source.

Carbon intensity reflects the weighted average of carbon emissions per unit of electricity generated, considering the contribution of each energy source. Adopting dynamic carbon intensity modeling allows for more accurate environmental assessments than using a fixed emission factor, facilitating the pursuit of net-zero emissions in buildings. Real-time CI data, provided by grid operators [6] or third-party services like Electricity Map [11] and WattTime [15], enhances the precision of carbon emission calculations, supporting informed decision-making in energy management.

Carbon capture and utilization: The CO_2 emission captured and utilized at time t can be indicated in (3.5). Part of the captured CO_2 can be used as raw material for P2G process, another portion can be stored, and the rest will be emitted into the atmosphere, which can be formulated in

$$E_{CCS,t}^{cap} = \eta_{CCS}(E_{CHP,t} + E_{GB,t}), \tag{3.5}$$

$$F_{g,t}^{P2G} = \eta_{P2G} P_{e,t}^{P2G}, \tag{3.6}$$

$$E_{CCS,t}^{P2G} = \eta_{CC} P_{e,t}^{P2G}, \tag{3.7}$$

$$E_{CCS,t}^{cs} = E_{CCS,t}^{cap} - E_{CCS,t}^{P2G} - E_{CCS,t}^{atm}, \tag{3.8}$$

where $E_{CSS,t}^{cap}$ represents the CO_2 captured via CCS at time t. $F_{g,t}^{P2G}$ is the natural gas produced via P2G process at time t. $E_{CCS,t}^{cs}$ and $E_{CCS,t}^{atm}$ represent the CO_2 is stored and emitted into the atmosphere, respectively. $P_{e,t}^{P2G}$ denotes the power consumption during the process of P2G.

In addition, the electricity consumption of CCS process at time t is indicated in

$$P_{e,t}^{CCS} = \eta_{cap} E_{CSS,t}^{cap}, \tag{3.9}$$

where $P_{e,t}^{CSS}$ is the power consumption during the process of CSS.

The overall direct carbon emission is determined by the difference between the emission resulting from natural gas consumption and the amount of CO_2 that is stored and utilized, which is illustrated by

$$\begin{aligned} E_{real} = & \sum\nolimits_{t=1}^{T} E_{CHP,t} + \sum\nolimits_{t=1}^{T} E_{GB,t} \\ & - \sum\nolimits_{t=1}^{T} \left(E_{CCS,t}^{P2G} + E_{CCS,t}^{cs} \right). \end{aligned} \tag{3.10}$$

The total carbon emission is the sum of direct real and indirect virtual carbon emissions, which is calculated by

$$E_{total} = E_{real} + E_{virtual}. \tag{3.11}$$

According to the carbon emission model, there are several strategies for the system to reduce its total carbon emissions. First, the system owner can directly decrease actual carbon emissions through the implementation of CCS technologies. Second, integrating zero-emission energy sources such as PV can substitute for energy purchases, thereby reducing reliance on non-renewable sources. Third, the system can minimize indirect carbon emissions from energy purchasing by shifting load consumption to periods when carbon intensity is lower.

3.3.2 Model of Building Integrated Energy System (BIES)

The other part of the model of building integrated energy system is shown in

$$Q_t^{ESS} = Q_{t-1}^{ESS} + P_{e,t}^{ESS}, \tag{3.12}$$

$$P_{min}^{ESS} \leq P_{e,t}^{ESS} \leq P_{max}^{ESS}, \tag{3.13}$$

$$Q_{min}^{ESS} \leq Q_t^{ESS} \leq Q_{max}^{ESS}, \tag{3.14}$$

$$P_{h,t}^{CHP} = \eta_{CHP,h} F_{CHP,t}, \tag{3.15}$$

$$P_{e,t}^{CHP} = \eta_{CHP,e} F_{CHP,t}, \tag{3.16}$$

$$P_{h,t}^{GB} = \eta_{GB} F_{GB,t}, \tag{3.17}$$

$$P_{h,t}^{AC} = \eta_{AC} P_{c,t}^{AC}, \tag{3.18}$$

$$P_{c,t}^{EC} = \eta_{EC} P_{e,t}^{EC}. \tag{3.19}$$

where $Q_{ess,t}$ represents the electricity storage level of the ESS at time t. $P_{ess,t}$ is charging and discharging power of ESS at time t (positive when charging and negative when discharging). $P_{h,t}^{CHP}$ and $P_{e,t}^{CHP}$ denote heat and power generation of CHP at time t. $P_{h,t}^{CHP}$ and $P_{e,t}^{CHP}$ are heat and power generation of CHP at time t, respectively. $P_{c,t}^{AC}$ and $P_{c,t}^{EC}$ represent the refrigeration power of AC and EC, respectively. $P_{h,t}^{AC}$ denotes heat absorption power at time t. $P_{e,t}^{EC}$ is electricity consumption power of EC at time t. We utilize subscripts e, h, and c to denote the load associated with electricity, heating, and cooling, respectively. Eq. (5.15) - Eq. (5.16) limits the operation of the energy storage system. Eq. (3.15) - Eq. (3.19) are the operation models of heat production and refrigeration equipment.

3.3.3 Chance Constraints

The modeling of the power balance for electricity, heating, and cooling, as well as the carbon consumption balance, is as follows:

1) The electricity load balance constraint can be written as

$$\begin{aligned}\Pr\Big(&P_{e,t}^{CHP}+P_{e,t}^{PV}+P_{e,t}^{buy}\geq P_{e,t}^{load}+P_{ess,t}\\&+P_{e,t}^{EC}+P_{e,t}^{P2G}+P_{e,t}^{CCS}\Big)\geq 1-\epsilon,\forall t\in\mathcal{T},\end{aligned}\tag{3.20}$$

where, $P_{e,t}^{\text{buy}}$ is the purchased electricity at time t; $P_{e,t}^{PV}$ is the PV generation at time t; $P_{e,t}^{load}$ is the load of electricity at time t. Eq. (5.18) describes the chance-constraint of power balance.

2) The heating load balance constraint can be written as

$$\Pr\left(P_{h,t}^{CHP}+P_{h,t}^{GB}\geq P_{h,t}^{load}+P_{h,t}^{AC}\right)\geq 1-\epsilon,\forall t\in\mathcal{T},\tag{3.21}$$

where $P_{h,t}^{load}$ is the heating load at time t. Eq. (3.21) describes the chance-constraint of thermal power balance.

3) The cooling load balance constraint is given by

$$\Pr\left(P_{c,t}^{EC}+P_{c,t}^{AC}\geq P_{c,t}^{load}\right)\geq 1-\epsilon,\forall t\in\mathcal{T},\tag{3.22}$$

where $P_{c,t}^{load}$ is the cooling load at time t. Eq. (3.22) describes the chance-constraint of cold power balance.

4) The carbon consumption balance constraint is given by

$$\Pr\left(E_{virtual}+E_{real}-E_{buy}\leq Q_{quota}\right)\geq 1-\epsilon,\tag{3.23}$$

where E_{buy} is the carbon allowance purchased from the carbon market, Q_{quota} is the original carbon quota the BIES held in time period T. The building owner can sell the surplus carbon allowance to the market to make profits. Equation (3.23) describes the chance constraint of carbon allowance balance, which is identified as the key constraint to achieving net-zero emission.

3.3.4 Objective Function

In order to explore the economic and environmental benefits of BIES operation, the total cost includes four terms, namely energy purchased cost (EPC), carbon storage cost (CSC), and carbon trading cost (CTC).

$$Cost = C_{EPC} + C_{CSC} + C_{CTC}. \tag{3.24}$$

Energy Purchased Cost (EPC): This cost incorporates the expenses for buying electricity and natural gas, represented by:

$$\begin{aligned} C_{EPC} = & \sum\nolimits_{t=1}^{T} p_{e,t} \cdot P_{e,t}^{buy} \\ & + \sum\nolimits_{t=1}^{T} p_{gas,t} \cdot (F_{CHP,t} + F_{GB,t} - F_{g,t}^{P2G}). \end{aligned} \tag{3.25}$$

Carbon Storage Cost (CSC): Costs associated with storing captured CO_2 are calculated as:

$$C_{CSC} = \sum\nolimits_{t=1}^{T} k_{cs} \cdot E_{CCS,t}^{cs} \tag{3.26}$$

where k_{cs} is the cost of carbon storage.

Carbon Trading Cost (CTC): This reflects the expenses or revenues from buying or selling carbon allowances in the carbon market:

$$C_{CTC} = p_{CO_2} \cdot E_{buy}, \tag{3.27}$$

where p_{CO_2} is the carbon price, and E_{buy} denotes the net carbon allowances traded.

To achieve net-zero emissions, it is assumed that building owners can purchase carbon allowances to offset their emissions or sell surplus allowances to make profits. The adoption of carbon capture and storage technologies, supported by a carbon market, offers a pathway to economically viable emission reductions. This approach enables BIES operators to balance emissions with carbon trading, optimizing both economic and environmental outcomes.

The chance-constrained (CC) form of the BIES operation problem is as follows:

$$\begin{aligned} \text{(CC)} \quad & \min C_{EPC} + C_{CSC} + C_{CTC} \\ & s.t. \quad Constraints\,(3.1) - (3.23) \end{aligned} \tag{3.28}$$

Eq. (5.18) - Eq. (3.23) are chance constraints. The target of the problem is to identify a feasible solution for (3.28) with a specified level of statistical confidence while minimizing the objective value as much as possible. In traditional CCP, knowing the distributions of the random variables is essential. With this knowledge, the non-convex CCs can be equivalently transformed into a tractable form. Detailed explanations of the transformations are provided in Appendix B.

3.4 Methodology

In this section, we begin by revisiting the formulation of chance-constrained programming (CCP) and present the fundamental procedural framework of our algorithm. Subsequently, we present the methodologies that have been developed in our research.

3.4.1 Proposed Framework

The common form of a CCP typically takes the following structure:

$$\min\ f(x);\ \text{s.t. } P(g(x;\xi) \in \mathcal{A}) \geq 1 - \epsilon \tag{3.29}$$

where x is the decision variable, ξ is a random variable that can be observed via a finite amount of data, $\mathcal{A}$ is the feasible region, and ϵ is the tolerance level. Prior to delving into the specifics, the concept of statistical feasibility is introduced.

Definition 3.4.1 (**Statistical Feasibility** [89]). *For a CC like (5.21), an algorithm is considered to be statistically feasible if the resulting solution, denoted as x^*, is feasible for the chance constraint with a confidence level of $1-\delta$ for any given dataset D_ξ.*

$$\mathbb{P}_{\mathcal{D}_\xi}\left(\mathbb{P}_\xi\left(g(x^*;\xi) \in \mathcal{A}\right) \geq 1-\epsilon\right) \geq 1-\delta \tag{3.30}$$

The inner constraint in (3.30) represents the conventional CC. Additionally, the outer constraint in (3.30) ensures that for all resulting solutions corresponding to different sample sets D_ξ, the inner chance constraint can be satisfied with a probability greater than or equal to $1-\delta$.

The original chance constraints are intractable, necessitating the use of an approximation. To tackle this challenging problem, we transform the model as outlined in (5.21) into a robust optimization (RO) problem. This transformation also facilitates the direct use of data [48]. RO addresses uncertainty by representing it through a deterministic set, referred to as an uncertainty set. The RO formulation ensures that the safety condition holds for any value of ξ within the uncertainty set. This helps to simplify the problem and make it more tractable. The RO form of (5.21) is as follows:

$$g(x;\xi) \in \mathcal{A}, \quad \forall \xi \in \mathcal{U} \tag{3.31}$$

where $\mathcal{U} \in \Omega$ represents an uncertainty set. It is evident that for any x that is feasible for (3.31), if $\xi \in \mathcal{U}$ then $g(x;\xi) \in \mathcal{A}$. Therefore, by selecting an uncertainty set $\mathcal{U}$ such that $P(\xi \in \mathcal{U}) \geq 1-\epsilon$, any feasible solution x for (3.31) must satisfy $P(g(x;\xi) \in \mathcal{A}) \geq P(\xi \in \mathcal{U}) \geq 1-\epsilon$.

The motivation behind this transformation is rooted in the fact that the RO form only necessitates an uncertainty set that approximately encompasses the potential values of ξ. As a result, there is no need to make assumptions about a specific distribution for ξ. This characteristic empowers the construction of uncertainty sets learning from sample data, making the approach free from distributional assumptions.

Various advanced forecasting algorithms can be applied to predict day-ahead load and PV generation [156]. Besides, novel approaches for forecasting carbon intensity have also emerged [120]. Nonetheless, these prediction models inevitably entail some level of prediction error. These uncertainties in the prediction process can have substantial consequences on energy system planning and operation. To account for these uncertainties, expressions for uncertainties associated with loads, PV generation, and carbon intensity are as follows:

$$P_{e,t}^{PV} = \overline{P}_{e,t}^{PV} + \tilde{P}_{e,t}^{PV}, \tag{3.32}$$

$$P_{e,t}^{\text{load}} = \overline{P}_{e,t}^{\text{load}} + \tilde{P}_{e,t}^{\text{load}}, \tag{3.33}$$

$$P_{h,t}^{\text{load}} = \overline{P}_{h,t}^{\text{load}} + \tilde{P}_{h,t}^{\text{load}}, \tag{3.34}$$

$$P_{c,t}^{\text{load}} = \overline{P}_{c,t}^{\text{load}} + \tilde{P}_{c,t}^{\text{load}}, \tag{3.35}$$

$$I_t = \overline{I}_t + \tilde{I}_t, \tag{3.36}$$

where $\overline{P}_{e,t}^{PV}, \overline{P}_{e,t}^{load}, \overline{P}_{h,t}^{load}, \overline{P}_{c,t}^{load}, \overline{I}_t$ are predicted value of day-ahead PV generation, electricity load, heating load, cooling load, and carbon intensity, respectively. The random terms are denoted by $\tilde{P}_{e,t}^{PV}, \tilde{P}_{e,t}^{load}, \tilde{P}_{h,t}^{load}, \tilde{P}_{c,t}^{load}, \tilde{I}_t$ and the probability distributions of random terms are not known.

Therefore, the constraints in (5.18) - (3.23) can be explicitly expressed in terms of random variables in the following form:

$$P_{e,t}^{CHP} + \overline{P}_{e,t}^{PV} + P_{e,t}^{buy} \geq \overline{P}_{e,t}^{load} + P_{e,t}^{ESS} + P_{e,t}^{EC} + P_{e,t}^{P2G}$$
$$+ P_{e,t}^{CCS} + \mathbf{h}^{\mathrm{T}}\boldsymbol{\xi}_{1,t}, \quad \forall \xi_{1,t} \in \mathcal{U}_{1,t}, \forall t \in \mathcal{T}, \tag{3.37}$$
$$P_{h,t}^{CHP} + P_{h,t}^{GB} \geq \overline{P}_{h,t}^{load} + P_{h,t}^{AC} + \xi_{2,t}, \quad \forall \xi_{2,t} \in \mathcal{U}_{2,t}, \forall t \in \mathcal{T}, \tag{3.38}$$
$$P_{c,t}^{EC} + P_{c,t}^{AC} \geq \overline{P}_{c,t}^{load} + \xi_{3,t}, \quad \forall \xi_{3,t} \in \mathcal{U}_{3,t}, \forall t \in \mathcal{T}, \tag{3.39}$$
$$\sum\nolimits_{t=1}^{T} (\bar{I}_t + \xi_{4,t}) * P_{buy,t} + E_{real} - E_{buy} \leq Q_{quota}, \forall \boldsymbol{\xi_4} \in \mathcal{U}_4, \tag{3.40}$$

where $\boldsymbol{\xi}_{1,t} = [\tilde{P}_{e,t}^{load}, \tilde{P}_{e,t}^{PV}]^\top$, $\mathbf{h} = [1, -1]^\top$,$\xi_{2,t} = \tilde{P}_{h,t}^{load}$, $\xi_{3,t} = \tilde{P}_{c,t}^{load}$, $\xi_{4,t} = \tilde{I}_t$. $\boldsymbol{\xi_4}$ is the vector form of $\xi_{4,t}, \forall t \in \mathcal{T}$. The learning-based robust optimization (LRO) problem is as follows:

$$\text{(LRO)} \quad \min C_{EPC} + C_{CSC} + C_{CTC}$$
$$s.t. \quad Constraints\, (3.1) - (3.19), (3.37) - (3.40). \tag{3.41}$$

3.4.2 Learning-based Robust Optimization (LRO)

Suppose the data set is $D = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$, $\boldsymbol{\xi_i} \in \mathbb{R}^m$, the question then revolves around how to construct the uncertainty set. The process is considered as learning the shape of the uncertainty set by utilizing historical data. The fundamental approach aims to shrink or reduce the size of the uncertainty set $\mathcal{U} = \mathcal{U}(D)$, which covers at least $1 - \epsilon$ of the true distribution P with a confidence level of $1 - \delta$, i.e.

$$\mathbb{P}_D(P(\boldsymbol{\xi} \in \mathcal{U}(D)) \geq 1 - \epsilon) \geq 1 - \delta. \tag{3.42}$$

To reduce the conservativeness of the solution, two principles are employed in the construction of the uncertainty set: 1) minimizing the set size as much as possible, and 2) ensuring that $P(\boldsymbol{\xi} \in \mathcal{U}(D))$ is not only greater than $1 - \epsilon$, but also approaches $1 - \delta$ closely.

Following the aforementioned principles, a two-phase strategy is adopted to construct the uncertainty set $\mathcal{U}$. The data D is partitioned into two groups, denoted as D_1 and

D_2, with sizes n_1 and n_2 respectively. Let $D_1 = \{\boldsymbol{\xi}_1^1, ..., \boldsymbol{\xi}_{n_1}^1\}$ and $D_2 = \{\boldsymbol{\xi}_1^2, ..., \boldsymbol{\xi}_{n_2}^2\}$. The two phases can be outlined as follows:

Phase 1: Shape learning. To approximate the shape of the uncertainty set, we utilize D_1 and employ an ellipsoidal uncertainty set, which is widely used in classical robust optimization to determine the shape of samples due to its tractability [47]. It is characterized by $\left\{\boldsymbol{\xi} \mid (\boldsymbol{\xi}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu}) \leq s\right\}$, where s is a positive constant. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the sample mean and covariance matrix of D_1, respectively. It enables us to determine the position of the smallest uncertainty set while providing a tractable approximation.

Phase 2: Shape calibration. To calibrate the size of the uncertainty set, we leverage D_2 and determine the appropriate value of s. The objective is to ensure that it satisfies (3.42) and that $\mathbb{P}_D(P(\boldsymbol{\xi} \in \mathcal{U}(D)) \approx 1-\epsilon)$ is close to $1-\delta$.

We first define a dimension-collapsing map $h(\cdot) : \mathbb{R}^m \to \mathbb{R}$ as follows:

$$h(\boldsymbol{\xi}) = (\boldsymbol{\xi}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu}). \tag{3.43}$$

Using the transformation map $h(\boldsymbol{\xi})$, each sample vector ξ can be mapped to a scalar value. We apply this transformation to all $\boldsymbol{\xi}_i^2 \in D_2$ and subsequently sort the resulting values $h(\boldsymbol{\xi}_i^2)$ in ascending order to get a new list denoted as $\{\hat{s}_k, 1 \leq k \leq n_2\}$. In other words, the values in the list satisfy the inequality $\hat{s}_k \leq \hat{s}_{k+1}$. The desired value of s corresponds to the ranking k^* of the sorted list $\{\hat{s}_k, 1 \leq k \leq n_2\}$. The index k^* is chosen such that:

$$k^* = \min\left\{r : \sum\nolimits_{k=0}^{r-1} C_{n_2}^k \left(1-\epsilon\right)^k \left(\epsilon\right)^{n_2-k} \geq 1-\delta\right\}. \tag{3.44}$$

Note that here we employ an ellipsoidal uncertainty set; however, it is also applicable to use a polyhedral uncertainty set as it is typical in the classical RO approach [47]. The procedure for utilizing the polyhedron uncertainty set remains the same. The process for constructing the uncertainty set is demonstrated in Algorithm 1 in detail. We have the following statistical guarantee of the constructed uncertainty set.

Theorem 3.4.1 (Statistical guarantee for LRO). *For the two datasets,* $D_1 = \{\boldsymbol{\xi}_1^1, ..., \boldsymbol{\xi}_{n_1}^1\}$ *and* $D_2 = \{\boldsymbol{\xi}_1^2, ..., \boldsymbol{\xi}_{n_2}^2\}$*. If* $n_2 \geq \log\delta/\log(1-\epsilon)$*, the obtained solution* $\hat{x}_0$ *of (5.21) using the construct uncertainty set* $\mathcal{U}$ *satisfies the statistical feasibility guarantee.*

More details of this proof can be found in [89]. Note that the sample requirement depends solely on δ and ϵ and is independent of the dimension of the random variable.

By transferring the CCs to their counterpart approximation using the classic RO form, these constraints can be subsequently converted into a tractable form once the uncertainty set has been constructed, as follows:

Theorem 3.4.2. *The constraints in (3.37) - (3.40) can be equivalently transformed into the following constraints:*

$$P_{e,t}^{CHP} + \overline{P}_{e,t}^{PV} + P_{e,t}^{buy} \geq \overline{P}_{e,t}^{load} + P_{e,t}^{ESS} + P_{e,t}^{EC} + P_{e,t}^{P2G} + P_{e,t}^{CCS} + \phi_{1,t}, \quad \forall t \in \mathcal{T}, \tag{3.45}$$

$$P_{h,t}^{CHP} + P_{h,t}^{GB} \geq \overline{P}_{h,t}^{load} + P_{h,t}^{AC} + \phi_{2,t}, \quad \forall t \in \mathcal{T}, \tag{3.46}$$

$$P_{c,t}^{EC} + P_{c,t}^{AC} \geq \overline{P}_{c,t}^{load} + \phi_{3,t}, \quad \forall t \in \mathcal{T}, \tag{3.47}$$

$$\sum_{t=1}^{T} \overline{I}_t * P_{buy,t} + E_{real} - E_{buy} + \phi_4 \leq Q_{quota}, \tag{3.48}$$

where $\phi_{1,t}, \phi_{2,t}, \phi_{3,t}$, and ϕ_4 satisfy

$$\phi_{1,t} = \mathbf{h}^\top \boldsymbol{\mu}_{1,t} + \sqrt{s_{1,t}} \left\| \boldsymbol{\Sigma}_{1,t}^{1/2} \mathbf{h} \right\|_2, \tag{3.49}$$

$$\phi_{2,t} = \mu_{2,t} + \sqrt{s_{2,t}} \left\| \Sigma_{2,t}^{1/2} \right\|_2, \tag{3.50}$$

$$\phi_{3,t} = \mu_{3,t} + \sqrt{s_{3,t}} \left\| \Sigma_{3,t}^{1/2} \right\|_2, \tag{3.51}$$

$$\phi_4 = \boldsymbol{P}_{buy}^\top \boldsymbol{\mu}_{1,t} + \sqrt{s_{4,t}} \left\| \boldsymbol{\Sigma}_4^{1/2} \boldsymbol{P}_{buy} \right\|_2. \tag{3.52}$$

$\boldsymbol{P}_{buy}$ denotes the vector form of $P_{buy,t}$. The details of the proof are provided in Appendix B. Note that then the problem is a second-order cone program (SOCP) problem which is convex and can be solved by standard solvers such as CPLEX and

Gurobi. By utilizing the shape parameters of the constructed uncertainty set, we can then reformulate the new problem (3.41) as an equivalent deterministic problem that is more tractable.

Algorithm 1 Learning-based Uncertainty Set Construction

Input: $\overline{P}_{e,t}^{PV}, P_{e,t}^{PV}, \overline{P}_{e,t}^{load}, P_{e,t}^{load}, \overline{P}_{h,t}^{load}, P_{h,t}^{load}, \overline{P}_{c,t}^{load}, P_{c,t}^{load}\ \overline{I}_t, I_t, \delta, \epsilon$

Output: The uncertainty set $\mathcal{U}_{1,t}(\xi_{1,t})$, $\mathcal{U}_{2,t}(\xi_{2,t})$, $\mathcal{U}_{3,t}(\xi_{3,t})$, $\mathcal{U}_4(\boldsymbol{\xi}_4)$

1: Compute the set of electricity load, heating load, cooling load, and carbon intensity prediction errors $D_{1,t}, D_{2,t}, D_{3,t}, D_4$ according to Eq. (4.13) - (3.35), respectively;

2: Divide $D_{1,t}$ into two groups $D_{1,t}^1$ and $D_{1,t}^2$, with sizes n_1 and n_2, respectively;

3: Estimate the sample mean $\mu_{1,t}$ and covariance matrix $\Sigma_{1,t}$ on $D_{1,t}^1$;

4: Construct the dimension-collapsing map $h(\cdot)$ according to (3.43) and calculate the transformed value $t(\boldsymbol{\xi}_{1,t})$ for all $\boldsymbol{\xi}_{\mathbf{1,t}} \in D_{1,t}^2$;

5: Sort all sort the resulting values $t(\boldsymbol{\xi}_{1,t})$ in ascending order to obtain the new list $\{\hat{s}_k, 1 \leq k \leq n_2\}$ and compute the index k^* using Eq. (3.44) for $\mathcal{U}_{1,t}$;

6: Set $s_{1,t}$ according to the obtained value corresponding to index k^* and get the uncertainty set for sample $D_{1,t}$ $\mathcal{U}_{1,t}(\boldsymbol{\xi}_{1,t}) = \{\boldsymbol{\xi}_{\mathbf{1,t}} | t(\boldsymbol{\xi}_{1,t}) \leq s_{1,t}\}$;

7: Construct the uncertainty set for $\mathcal{U}_{2,t}(\xi_{2,t})$, $\mathcal{U}_{3,t}(\xi_{3,t})$, and $\mathcal{U}_4(\boldsymbol{\xi}_4)$ following the same process outlined in Steps 2 - 6.

3.4.3 Reconstructed Learning-based Robust Optimization

Despite constructing the uncertainty set in Section 3.4.2, the resulting solutions to the problem can still exhibit conservativeness. This is because the construction of $\mathcal{U}$ is independent of the obtained solution $\hat{x}$. Although $P(g(\hat{x}; \boldsymbol{\xi}) \in \mathcal{A}) \geq P(\boldsymbol{\xi} \in \mathcal{U})$, the actual confidence level of $\mathbb{P}_D(P(\boldsymbol{\xi} \in \mathcal{U}(D)) \geq 1 - \epsilon)$ remains conservative. In this section, our objective is to further mitigate the conservation present in the solutions.

The key idea in this approach is to incorporate the knowledge relevant to the optimization problem when constructing the uncertainty set. We can still use a two-phase approach to solve this problem. In Phase 1, the method described in Section 3.4.2 is utilized to identify the uncertainty set, and a solution $\hat{x}$ for (5.21) is obtained. The resulting uncertainty set takes the form $\{\xi | g(\hat{x_0}; \boldsymbol{\xi}) \in \mathcal{A}\}$. In Phase 2, the size of the reshaped set is adjusted, resulting in $\{\xi | g(\hat{x_0}; \boldsymbol{\xi}) \in \tilde{\mathcal{A}}\}$, where $\tilde{\mathcal{A}}$ represents the size-tuned uncertainty set.

In particular, we concentrate on analyzing the carbon balance condition (3.40), while similar analysis can be conducted for the other constraints. For an uncertainty set $\mathcal{U}_4$ and its corresponding solution $\{\hat{P}_{buy,t}, \hat{E}_{buy}, \hat{E}_{real}\}$, the following condition is satisfied:

$$\begin{aligned} &\mathbb{P}_{\xi}\left(\boldsymbol{\xi} \in \mathcal{U}_4\right) \leq \\ &\mathbb{P}_{\xi}\left(\sum_{t=1}^{T} (\overline{I}_t + \xi_{4,t}) * \hat{P}_{buy,t} + \hat{E}_{real} - \hat{E}_{buy} \leq Q_{quota}\right) \end{aligned} \tag{3.53}$$

Therefore, the following uncertainty set can be considered as a viable solution:

$$\mathcal{U}_4 = \left\{\boldsymbol{\xi} | \sum_{t=1}^{T} (\overline{I}_t + \xi_{4,t}) * \hat{P}_{buy,t} + \hat{E}_{real} - \hat{E}_{buy} \leq Q_{quota} + \alpha \right\} \tag{3.54}$$

Once again, we partition the dataset of carbon intensity forecast error D_4 into D_4^1 and D_4^2, with respective sizes n_1 and n_2. The dataset D_4^1 is used to estimate $\{\hat{P}_{buy,t}, \hat{E}_{buy}, \hat{E}_{real}\}$, while D_4^2 is used to estimate α. In this section, we present the construction process for $\mathcal{U}_4'$, while noting that the construction of $\mathcal{U}_{1,t}', \mathcal{U}_{2,t}'$, and $\mathcal{U}_{3,t}'$ can be accomplished using the same procedure. The two-phase approach is outlined as follows:

Phase 1: Shape learning. Run algorithm 1 based on D_4^1 to get the initial solution $\{\tilde{P}_{buy,t}, \tilde{E}_{buy}, \tilde{E}_{real}\}$.

Phase 2: Shape calibration. The calibration for α follows a similar approach as that in LRO. To achieve this, we introduce a customized dimension-collapsing map $h_{Recon}(\cdot) : \mathbb{R}^m \to \mathbb{R}$, which is designed as follows:

$$h_{Recon}(\boldsymbol{\xi}) = \sum_{t=1}^{T} (\overline{I}_t + \xi_{4,t}) * \hat{P}_{buy,t} + \hat{E}_{real} - \hat{E}_{buy} - Q_{quota} \tag{3.55}$$

Following a similar approach as in LRO, we employ the transformation $h_{Recon}(\boldsymbol{\xi})$ to map each sample vector $\boldsymbol{\xi}_i$ in D_1 to a scalar value, which is then sorted in ascending order. The desired value corresponds to the ranking k^*, where k^* is computed using (3.44). The process for constructing the uncertainty set is detailed in Algorithm 2. And the performance improvement guarantee indicates that by reconstructing the uncertainty set, despite the slightly increased computational cost, we can enhance the performance of the original LRO.

Algorithm 2 Uncertainty Set Reconstruction

Input: $\overline{P}_{e,t}^{PV}, P_{e,t}^{PV}, \overline{P}_{e,t}^{load}, P_{e,t}^{load}, \overline{P}_{h,t}^{load}, P_{h,t}^{load}, \overline{P}_{c,t}^{load}, P_{c,t}^{load}\ \overline{I}_t, I_t, \delta, \epsilon$

Output: The reconstructed uncertainty set $\mathcal{U}_{1,t}(\boldsymbol{\xi}_{1,t})'$, $\mathcal{U}_{2,t}(\xi_{2,t})'$, $\mathcal{U}_{3,t}(\xi_{3,t})'$, $\mathcal{U}_4(\boldsymbol{\xi}_4)'$

1: Compute the set of electricity load, heating load, cooling load, and carbon intensity prediction errors $D_{1,t}, D_{2,t}, D_{3,t}, D_4$ according to (4.13) - (3.35), respectively;

2: Divide D_4 into two groups D_4^1 and D_4^2, with sizes n_1 and n_2, respectively;

3: Call Algorithm 1 to compute the approximate solutions $\{\hat{P}_{buy,t}, \hat{E}_{buy}, \hat{E}_{real}\}$ with respect to D_4^1

4: Construct the dimension-collapsing map $h_{Recon}(\cdot)$ according to Eq.(3.55) and calculate the transformed value $h_{Recon}(\boldsymbol{\xi}_4)$ for all $\boldsymbol{\xi_4} \in D_4^2$;

5: Sort all sort the resulting values $h_{Recon}(\boldsymbol{\xi}_4)$ in ascending order to obtain the new list $\{\hat{s}_k, 1 \leq k \leq n_2\}$ and compute the index k^* using (3.44) for $\mathcal{U}_4'$;

6: Set α according to the obtained value corresponding to index k^* and get the uncertainty set for sample D_4 such that $\mathcal{U}_4(\boldsymbol{\xi}_4)' = \{\boldsymbol{\xi_4} | t(\boldsymbol{\xi}_4) \leq \alpha\}$;

7: Construct the uncertainty set for $\mathcal{U}_{1,t}(\xi_{1,t})$, $\mathcal{U}_{2,t}(\xi_{2,t})$, and $\mathcal{U}_{3,t}(\xi_{3,t})$ following the same process outlined in Steps 2 - 6.

3.5 Numerical studies

3.5.1 Data Description and Experimental Settings

In this section, we conduct comprehensive numerical studies to demonstrate the accuracy of our carbon model, validate the performance of our proposed learning-based robust algorithms, and analyze the sensitivity of key parameters. The parameters required for the components in the BIES are listed in Table 1, which are set based on the research papers [72] and [176]. The time horizon of the delivery day is divided into 24 hours. For each setting, we repeat the experiments 300 times. All of the following simulations were implemented using Matlab R2022b software on a desktop computer with an Intel i9 2.4GHz CPU and 32GB RAM. The optimization problem is solved using CPLEX 12.8.

The electricity price data is sourced from [122], ensuring that our simulations are based on realistic market conditions. The building load and PV generation data are modified based on the End-Use Load Profiles for the U.S. Building Stock [161], a widely recognized dataset in building energy research. In Figure 3.2, the purple line represents the PV generation, while the blue, red, and yellow lines depict the electricity, heating, and cooling load profiles for a single day, respectively. The historical source mix data is obtained from California Independent System Operator [6]. Figure 3.3 illustrates the carbon intensity profile during a single day, reflecting the variations in carbon emissions throughout the day.

By utilizing realistic data and settings, we aim to ensure the validity and practicality of our simulations, enabling meaningful insights into the performance and sensitivity of the proposed learning-based robust algorithms.

3.5.2 Benchmarks and Evaluation Metrics

Our proposed methods as outlined in Sections III and IV are designated as LRO and ReLRO. In order to evaluate the performance of our approaches, we compare them with other common methods in decision-making under uncertainty:

1) *Chance-constrained Optimization (CC)*: The conventional chance-constrained methods [142], as commonly employed in various fields, typically operate under the assumption of a specific distribution family for the random variable. In this study, we make

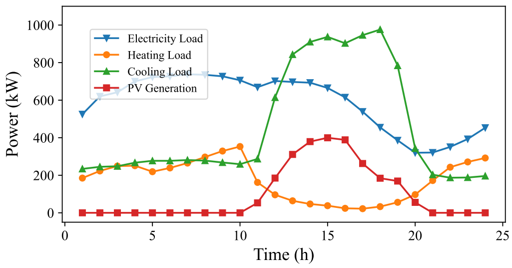


Figure 3.2: Load profile of the BIES

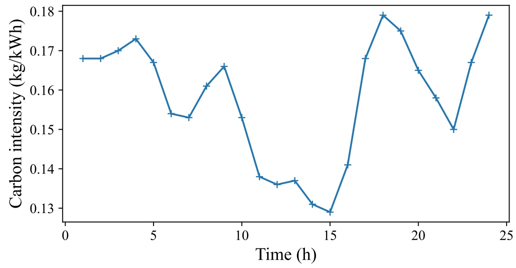


Figure 3.3: Day-ahead carbon intensity

the supposition that the prediction error conforms to a Gaussian distribution.

2) *Scenario Generation (SG)*: The scenario generation method, introduced in [56], replaces the chance constraint with a set of sampled constraints. It is important to note that SG relies on an adequate number of samples, and the running time of SG increases linearly with the size of the sample.

3) *Distributionally Robust Chance-Constrained (DRCC)*: Our proposed methods are compared with distributionally robust chance-constrained methods outlined in [57]. In this method, the probability distribution of the data is not fully required, but rather it is known to belong to a predetermined class of distributions.

4) *The Optimal Baseline (OPT)*: The optimal baseline method considered in this study is deterministic optimization with perfect predictions, denoted as OPT. In this approach, all predictions and input parameters are assumed to be completely accurate, allowing for an optimal solution to be determined without considering any uncertainty or variability.

For synthetic scenarios, we adopt the definition proposed in [89]. We denote $\hat{\epsilon}$ as the estimated expected violation probability of the obtained solution. More specifically, $\hat{\epsilon}$ is calculated as $\hat{E}_D\left[P_{\text{vio}}\right]$, where $\hat{E}_D[\cdot]$ represents the empirical expectation. Here, P_{vio} refers to the probability $P(g(\hat{x}(D);\xi) \notin \mathcal{A})$. Additionally, the statistic violation rate $\hat{\delta}$ is $\hat{P}_D\left(P_{\text{vio}} > \epsilon\right)$. For approaches not reliant on data, the chance constraint is always fulfilled, yielding $\hat{\delta} = 0$. The experiment is conducted on 300 different sample sets. We generate the samples randomly from a multivariate Gaussian distribution with a positive-definite covariance matrix, which is widely adopted to model the load prediction error in literature [80] [39]. Note that in the real-world data experiments, we do not assume the distribution and use the real prediction error. Since there are many chance constraints, we only focus on the carbon balance violation rate.

3.5.3 Performance Analysis for Synthetic Data

1) Sensitivity Analysis for Sample Size: To assess the influence of sample size, we conduct 300 trials for each of the five different approaches, varying the sample sizes. Figures 3.4 and 3.5 illustrate how the total cost and violation rate of different approaches change with the sample size. It is noted that the performance trend varies significantly among the different methods analyzed. Specifically, the total cost and violation rate of the CC and DRCC approaches do not change substantially with increasing sample size. This can be attributed to the fact that these methods primarily rely on parametric estimation of the sample distribution. Table 3.2 indicates that for the CC approach, the observed violation rate exceeds the prescribed requirement, with a statistic violation rate of approximately 0.5. For the SG method, the total cost increases with the rising sample size, while the violation rate decreases.

For our proposed ReLRO approach, Table 3.2 demonstrates that the statistic violation rate is approximately 0.05. With an increasing sample size, our ReLRO method effectively leverages the flexibility reserve provided by the chance constraints, providing a less conservative solution that satisfies the prescribed conditions while also exhibiting an increasing violation rate.

Table 3.3 presents the running time of the various approaches analyzed. For the SG approach, it is evident that the running time increases proportionally with larger sample sizes, which can make the computation time quite extensive for large sample sizes. However, the running time of the other algorithms shows minimal fluctuations

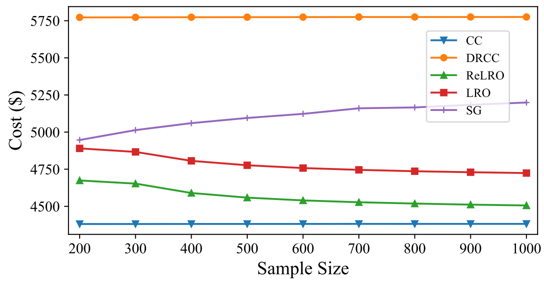


Figure 3.4: Cost under different sample sizes

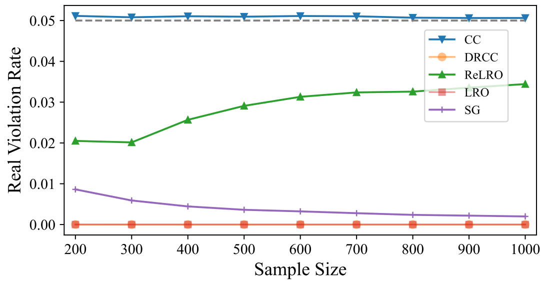


Figure 3.5: Real violation rate under different sample sizes

Table 3.2: Statistic violation rate under different sample sizes

Sample	CC	SG	DRCC	LRO	**ReLRO**
200	0.51	0	0	0	0.046667
300	0.446667	0	0	0	0.023333
400	0.503333	0	0	0	0.03
500	0.486667	0	0	0	0.033333
600	0.486667	0	0	0	0.056667
700	0.49	0	0	0	0.046667
800	0.48	0	0	0	0.053333
900	0.493333	0	0	0	0.046667
1000	0.486667	0	0	0	0.063333

with changing sample sizes, suggesting that the sample size does not noticeably impact the computation time for these methods. The running time of the ReLRO approach is slightly higher than the LRO because ReLRO can be regarded as solving the LRO twice. Despite the marginally increased runtime, ReLRO improves the performance of the solution, rendering the associated computational cost worthwhile.

2) Sensitivity Analysis for Violation Level Control: Figures 3.6 and 3.7 depict the performance of various approaches under different stability requirements $1-\epsilon$. First, we observe that for CC, the curve lies slightly above the stability requirement, which indicates that the CC approach does not guarantee compliance with the chance constraints in all cases. In contrast, the SG approach consistently generates feasible

Table 3.3: Running time under different sample sizes (s)

Sample	CC	SG	DRCC	LRO	**ReLRO**
200	1.136196	11.97335	1.173524	0.570546	2.705187
300	1.150944	13.76358	1.177605	0.588166	1.939536
400	1.142088	20.03871	1.433913	0.590878	1.768694
500	1.37538	26.45544	1.315433	0.584426	2.168233
600	1.15595	35.35467	1.164011	0.704867	2.232307
700	1.168317	40.99384	1.394985	0.58639	2.330702
800	1.678593	51.29287	1.158964	0.642918	1.993813
900	1.298453	57.48193	1.646797	0.69297	1.682798
1000	1.165364	67.98769	1.404033	0.643389	2.035873

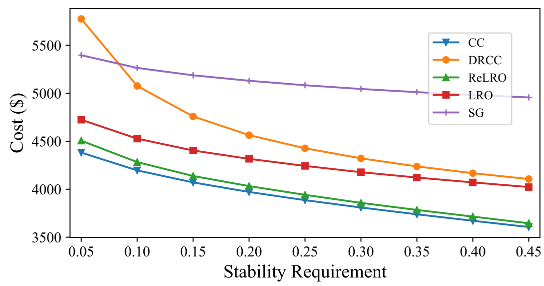


Figure 3.6: Cost of different stability requirements

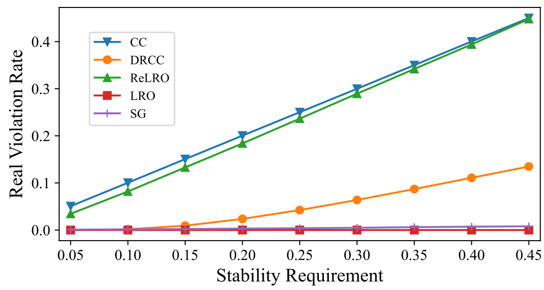


Figure 3.7: Real violation rate under different stability requirements

solutions while offering improved economic performance as $1 - \epsilon$ decreases. But the total cost is the highest when ϵ is greater than 0.1, which means that it is more conservative than other methods. However, the violation rate curve of our proposed ReLRO method consistently remains lower than the allowed risk. These findings emphasize the strength of our proposed ReLRO approach in terms of effectively leveraging flexibility reserves and providing solutions that are both cost-efficient and compliant with the given chance constraints, even under more stringent stability requirements.

3.5.4 Performance Analysis for Real Data

In this subsection, we evaluate the performance of our algorithm using real prediction data. To accomplish this, we construct neural networks dedicated to forecasting the

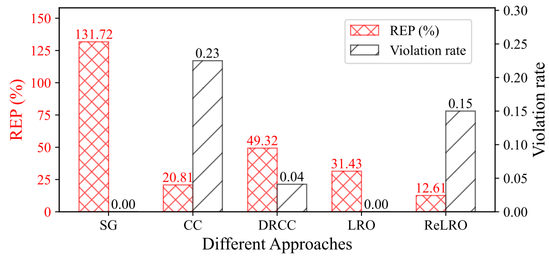


Figure 3.8: REP under different approaches ($\delta = 0.2$, $\xi = 0.2$)

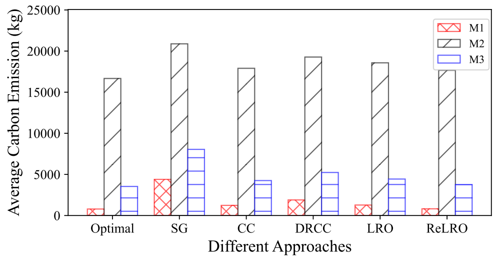


Figure 3.9: Average cost under different grid emission accounting models

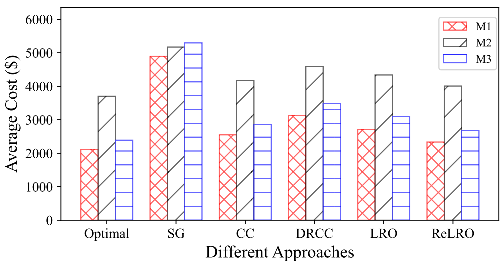


Figure 3.10: Average cost under different grid emission accounting models

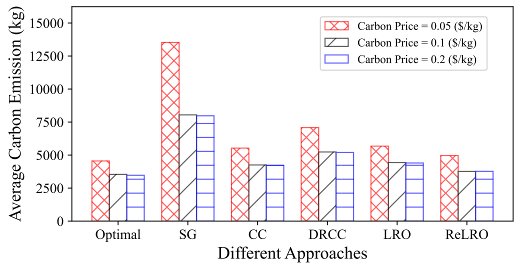


Figure 3.11: Carbon emission under different carbon prices

day-ahead load, PV generation, and carbon intensity. In order to make predictions, we utilize the data from the previous day. The training set is employed for both training the prediction model and constructing the uncertainty set, while the test set is utilized to calibrate the uncertainty set and evaluate the performance. And the result in this subsection is the average value in of 120 test days. *1) The performance of in real data.* In our evaluation, we quantify the average savings in the total cost by comparing it to the optimal benchmark OPT. This comparison is expressed as a percentage, donated as the regret percentage (REP). The REP represents the gap between the achieved cost and the optimal cost, thereby indicating the relative cost-effectiveness of the evaluated methods.

Figure 3.8 illustrates the average REP and violation rate. Notably, the ReLRO method achieves the lowest REP (12.67%), outperforming other methods by at least 8.2% while keeping the violation rate beneath the predetermined threshold of 0.2.

The findings from the real data experiment align with those observed in the synthetic dataset. The CC method exhibits a higher average REP (20.81%) than ReLRO and its violation rate surpasses the predefined threshold. It arises from the CC method's assumption of a Gaussian distribution in the data, which proves inadequate when the prediction errors deviate from this distribution. Despite the SG method achieving a zero violation rate, its REP is the highest, indicating an overconservative approach in the real dataset.

2) The impacts of different carbon emission accounting models: In this part, we evaluate three distinct scenarios, each employing a different model for carbon emission accounting. These models are designated as M1, M2, and M3 for ease of reference.

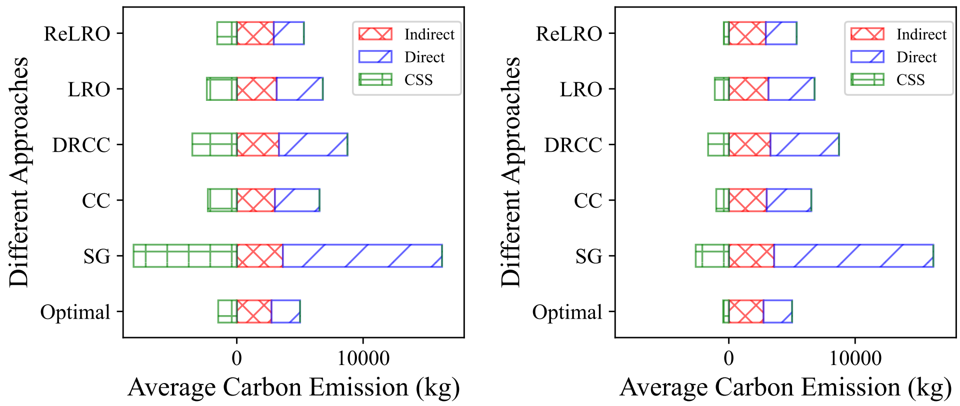
- M1: The emission factor is set at zero, indicating that we do not consider the carbon emission by the electricity purchased from the grid. It is widely employed in the literature [108] [124].
- M2: It is assumed that all electricity generation is coal-based, which is the same as the assumptions made in [176] and [178]. The emission intensity is set to 0.993 kg/kWh based on the report of [11].
- M3: This is the proposed real-time carbon intensity model. It calculates grid emissions on an hourly basis, offering a more accurate method for carbon accounting.

Figure 3.9 and 3.10 represent the average carbon emissions and costs under different carbon accounting models, respectively. With the grid emission factor at zero, as in M1, the model overlooks grid emissions, leading to a lower estimation of overall carbon emissions. This underestimation is critical as grid-sourced electricity is a major contributor to total emissions. In the M2 scenario, where the emission factor is 0.993 kg/kWh, assuming all electricity is coal-fired, there is a tendency to overestimate emissions. This overestimation occurs because the actual grid mix includes renewable energy sources. In contrast, our M3 model, which integrates real-time carbon intensity, provides a more nuanced and accurate depiction of the carbon system within BIES. This approach enhances the effectiveness of decision-making processes by offering a more realistic view of carbon emissions. Additionally, we believe that a key advantage of an accurate carbon accounting model is its potential to enable buildings to lower emissions through strategic load shifting, such as adjusting cooling or heating schedules and optimizing ESS operations. If the emission factor remains constant, buildings cannot achieve emission reductions by utilizing electricity during periods of low carbon intensity.

Besides, Figures 3.9 and 3.10 demonstrate that our methods effectively handle uncertainty and consistently outperform other methods under different carbon accounting models. In models M1 and M2, carbon emissions are accounted for using a fixed constant, thereby eliminating uncertainties related to carbon intensity. Across the various models, the SG method and the DRCC method tend to yield more conservative results, a finding that corroborates previous observations. In M3, our method stands out for its superior performance in handling different carbon accounting scenarios to reduce carbon emissions by at least 13.8%. But in M1, our model only achieves 1.4%. Because the total carbon emission is overestimated thus diminishing the impacts of uncertainties.

3) The impacts of different carbon prices: Figure 3.11 depicts the impact of different carbon pricing on emissions reduction through various methods. With perfect prediction, comparing a carbon price increase from \$0.05/kg to \$0.10/kg results in a 22.36% decrease in average emissions. Yet, elevating the price to \$0.20/kg only reduces emissions by 1.90%, indicating diminishing returns at higher price levels due to limits in the system's adaptability. These findings reveal a significant relationship between increasing carbon prices and decreased carbon emissions, emphasizing the effectiveness of appropriate carbon pricing strategies in reducing carbon footprint.

4) The effectiveness of CCS: In this subsection, we assess the effectiveness of CSS



(a) $p_{co_2} = \$0.10/\text{kg}$ (b) $p_{co_2} = \$0.05/\text{kg}$

Figure 3.12: Direct emission, indirect emission and carbon reduction by CSS under different carbon prices

technology. It should be noted that implementing CCS technology as a standalone solution for emission reduction may not be economically feasible. The integration of a carbon market could provide an incentive for companies to adopt CCS technology by converting emission reductions into economic advantages. Figure 3.12 illustrates the direct emissions, indirect emissions, and carbon captured by the CSS under varying carbon price scenarios. In the optimal scenario where the carbon price is \$0.10/kg, the CCS system reduces total carbon emissions by 29.35%. However, at a lower carbon price of \$0.05/kg, the reduction in carbon emissions drops to 9.01%. This underscores that the carbon price is a crucial determinant of CCS utilization.

3.6 Conclusion

Decarbonizing the building sector is crucial for achieving climate neutrality. In this paper, we have devised a carbon-conscious optimization strategy to enable the efficient operation of diverse components and processes within building integrated energy systems, such as combined heat and power, energy storage, carbon capture, and the import of electricity from the primary grid. The proposed model aims to minimize the overall energy expenses while simultaneously curbing carbon emissions toward net-zero. Furthermore, to address the uncertainties inherent in the BIES, we have developed a robust learning-based optimization framework that provides statistical feasibility guarantees. The approach effectively handles the uncertainties and ensures reliable and robust operation by designing the appropriate uncertainty set. The proposed approach has been validated through numerical studies conducted using both synthetic and real data. Using synthetic data, our approach demonstrates improved control over the violation rate and leads to lower costs. Meanwhile, when applied to real-world data, the proposed approach achieves a significant reduction in regret percentage of over 8% for BIES control, compared to other methods in the literature. The algorithms in this work use an ellipsoidal uncertainty set to learn the shape of uncertainties. Future research could investigate the application of deep learning techniques to learning a more complex shape of the uncertainty set, thereby enhancing the precision.

3.7 Appendix

3.7.1 Details about the Prediction Models

In the numerical studies, we construct distinct neural networks to predict day-ahead load, photovoltaic power generation, and carbon intensity, respectively. We employ the widely used multi-layer perceptron (MLP) [99] models for predictions. Considering the example of electricity load prediction, the model utilizes demand data from the preceding 24 hours as input to forecast the subsequent day's demand. Additional parameters are detailed in Table 3.4. We partition the dataset into two segments: 55% for training the prediction model and constructing uncertainty sets, and 45% for testing and calibrating the uncertainty set.

Table 3.4: Relevant parameters for prediction models

Parameters	Values
Hidden layer size	256
# of hidden layers	2
Batch size	16
Optimizer	Adam
Activation	ReLU
Learning rate	0.001

3.7.2 Reformulation for Traditional CCs

For the traditional CCP, it is necessary to know the distribution of the random variable. Here, we use constraint (5.18) as an example. If $\boldsymbol{\xi_{1,t}}$ follows Gaussian distribution, the intractable CC can be transformed into a deterministic convex problem, as outlined in [142].

Consider $\boldsymbol{\xi}_{1,t} \sim N\left(\boldsymbol{\mu}_{1,t}, \boldsymbol{\Sigma}_{1,t}\right)$, and Φ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution. It follows that

$$\operatorname{Prob}\left(\mathbf{h}^{\top}\boldsymbol{\xi}_{1,t} \leq b\right) = \Phi\left(\frac{b - \mathbf{h}^{\top}\boldsymbol{\mu}_{1,t}}{\sqrt{\mathbf{h}^{\top}\boldsymbol{\Sigma}_{1,t}\mathbf{h}}}\right), \tag{3.56}$$

where

$$b = \overline{P}_{e,t}^{load} + P_{e,t}^{ESS} + P_{e,t}^{EC} + P_{e,t}^{P2G} - (P_{e,t}^{CHP} + \overline{P}_{e,t}^{PV} + P_{e,t}^{buy}). \tag{3.57}$$

Thus,

$$\begin{aligned} &\operatorname{Prob}\left(\mathbf{h}^{\top}\boldsymbol{\xi}_{1,t} \leq b\right) \geq 1 - \epsilon \Longleftrightarrow \\ &b - \mathbf{h}^{\top}\boldsymbol{\mu}_{1,t} \geq \Phi^{-1}(1-\epsilon)\left\|\boldsymbol{\Sigma}_{1,t}^{1/2}\mathbf{h}\right\|_2 \end{aligned} \tag{3.58}$$

Therefore, the equivalent constraints are:

$$\begin{aligned} &P_{e,t}^{CHP} + \overline{P}_{e,t}^{PV} + P_{e,t}^{buy} \geq \overline{P}_{e,t}^{load} + P_{e,t}^{ESS} + P_{e,t}^{EC} + P_{e,t}^{P2G} \\ &+ P_{e,t}^{CCS} + \mathbf{h}^{\top}\boldsymbol{\mu}_{1,t} + \Phi^{-1}(1-\epsilon)\left\|\boldsymbol{\Sigma}_{1,t}^{1/2}\mathbf{h}\right\|_2 \end{aligned} \tag{3.59}$$

By following this procedure, deterministic formulations for the remaining chance constraints can be derived. Note that when the random variable has an additive relationship with the decision variables, it results in linear constraints. When the random variable has a multiplicative relationship with the decision variables, it results in second-order cone constraints. Both types of constraints can be solved efficiently using standard solvers.

3.7.3 Proof for Theorem 2

By utilizing the shape parameters of the constructed uncertainty set, the problem (3.41) can be reformulated as a more tractable equivalent one. As an example, consider constraint (3.37) and the constructed ellipsoidal uncertainty set as follows:

$$\mathcal{U}_{1,t} = \left\{ \boldsymbol{\xi}_{1,t} : \left(\boldsymbol{\xi}_{1,t} - \boldsymbol{\mu}_{1,t}\right)^\top \boldsymbol{\Sigma}_{1,t}^{-1} \left(\boldsymbol{\omega}_{1,t} - \boldsymbol{\mu}_{1,t}\right) \leq s_{1,t} \right\}, \tag{3.60}$$

where $\boldsymbol{\mu}_{1,t}, \boldsymbol{\Sigma}_{1,t}$ represent the sample mean and covariance matrix of $\boldsymbol{\xi}_{1,t}$, respectively, and $s_{1,t}$ is a positive scalar value determined using Algorithm 1. The RO form is:

$$\begin{aligned} &P_{e,t}^{CHP} + \overline{P}_{e,t}^{PV} + P_{e,t}^{buy} \geq \overline{P}_{e,t}^{load} + P_{e,t}^{ESS} + P_{e,t}^{EC} + P_{e,t}^{P2G} \\ &+ P_{e,t}^{CCS} + \mathbf{h}^\top \boldsymbol{\xi}_{1,t}, \quad \forall \xi_{1,t} \in \mathcal{U}_{1,t}, \forall t \in \mathcal{T}. \end{aligned} \tag{3.61}$$

Note that the random variable has an additive relationship with the decision variables. Thus, the following substitution can be made:

$$\begin{aligned} &P_{e,t}^{CHP} + \overline{P}_{e,t}^{PV} + P_{e,t}^{buy} \geq \overline{P}_{e,t}^{load} + P_{e,t}^{ESS} + P_{e,t}^{EC} + P_{e,t}^{P2G} \\ &+ P_{e,t}^{CCS} + \phi_{1,t}, \quad \forall t \in \mathcal{T} \end{aligned} \tag{3.62}$$

where $\phi_{1,t}$ satisfies:

$$\begin{aligned} \phi_{1,t} = \max \quad & \mathbf{h}^\top \boldsymbol{\xi}_{1,t} \\ \text{s.t. } & (\boldsymbol{\xi}_{1,t} - \boldsymbol{\mu}_{1,t})^\top \boldsymbol{\Sigma}_{1,t}^{-1} (\boldsymbol{\xi}_{1,t} - \boldsymbol{\mu}_{1,t}) \leq s_{1,t} \end{aligned} \tag{3.63}$$

Slater's condition ensures that strong duality holds, thus we have:

$$\begin{aligned}\phi_{1,t} = &\max_{\lambda} \min_{\boldsymbol{\xi}_{1,t}} -\mathbf{h}^\top \boldsymbol{\xi}_{1,t} \\ &+ \lambda \left((\boldsymbol{\xi}_{1,t} - \boldsymbol{\mu}_{1,t})^\top \boldsymbol{\Sigma}_{1,t}^{-1} (\boldsymbol{\xi}_{1,t} - \boldsymbol{\mu}_{1,t}) - s_{1,t} \right) \\ &\text{s.t.} \quad \lambda \geq 0 \end{aligned} \tag{3.64}$$

Due to the positive definiteness of the covariance matrix $\Sigma_{1,t}$, the quadratic programming problem can be solved directly. Consequently, we can also derive explicit representations for $\phi_{1,t}$, $\phi_{2,t}$, $\phi_{3,t}$, and ϕ_4, respectively.

Chapter 4

Carbon-aware Scheduling of Thermostatically Controlled Loads

4.1 Introduction

Thermostatically controllable loads (TCLs), such as refrigerators, heating, and ventilation and air conditioning (HVAC), are playing an ever-increasingly important role in consumer-level demand response programs [186]. TCLs encompass a range of controllable appliances equipped with thermal storage features, enabling them to be regulated according to the exchange of heat between outdoor temperatures and indoor air and maintain comfortable indoor conditions tailored to user preferences in residential and commercial buildings. Compared with conventional loads, TCLs' unique operational thermodynamics and the thermal inertia of buildings enable the storage of thermal energy within the building envelope for pre-cooling/heating purposes. Therefore, TCLs' power consumption can be adjusted temporally with a mild impact on consumers' thermal comfort, making them an ideal source for demand-side flexibility. Prior literature, such as [34] [69] [74], have demonstrated that TCLs have a crucial impact on the functioning of the electric grid by adjusting their electricity consumption during peak periods, which can lead to significant cost reductions while maintaining the comfort of building occupants.

On the other hand, TCLs also account for a significant share (e.g., 40% to 60%) of the total energy consumption within buildings [71]. Considering that the building sector constitutes 30% of global final energy usage and is responsible for 26% of global energy-related emissions, there is immense pressure on this sector to pioneer carbon-conscious decarbonization solutions and stay on track of the Net Zero Emissions by 2050 Scenario [163]. In this context, TCLs emerge as a critical component of the solution. Their ability to regulate energy consumption based on demand and optimize thermal storage presents an opportunity to significantly reduce carbon footprints in buildings.

While the operation and control of TCLs have been studied in the literature, two important gaps remain: the first gap lies in the lack of incorporation of a "carbon perspective" in TCL scheduling and control strategies. A majority of the prior studies focus on cost-minimization objectives or timely ancillary service provision when TCLs are involved in demand response. Given the urgency of decarbonizing commercial and residential buildings, new carbon-oriented operation strategies and algorithms need to be developed for TCL scheduling that allow decision-makers to define and plan the necessary short-term actions toward the goal of achieving zero-carbon-ready, as well as to track progress over time.

Another research challenge stems from uncertainties introduced by forecast errors in ambient temperature and base load predictions. The scheduling of TCLs typically relies on forecasts of outdoor temperature [85]. Forecast errors represent a significant factor influencing indoor temperature and may lead to deviations from the desired comfortable temperature range and violate occupant comfort requirement [160]. These errors also impact the building's interactions with the electrical grid, as additional power may need to be imported to compensate for the forecasting deviations. Therefore, scheduling TCLs must carefully consider the stochasticity of ambient temperature and be robust to inaccuracies in predictions [164] [171]. These uncertainties must be systematically and comprehensively incorporated into the decision-making process associated with TCL control to ensure the feasibility of building energy management across diverse operational scenarios without disrupting occupant comfort preferences.

Motivated by the aforementioned gaps, this paper presents a data-driven approach for low-/zero-emission TCL scheduling. To accurately account for carbon emissions, the proposed approach incorporates the time-varying carbon intensity of the regional grid within the TCL scheduling problem formulation, facilitating well-informed decision-making. To address the uncertainties stemming from outdoor temperature and base load (i.e., energy consumption excluding TCLs) forecast errors, we formulate the TCLs scheduling problem into a DRCC problem under Wasserstein ambiguity set. To overcome the conservatism of the CVaR approximation and computational intractability of the exact MILP approaches, we propose a novel reformulation strategy, drawing inspiration from the ALSO-X method proposed in [94], which recasts the DRCC into a corresponding bilevel problem. In this bi-level structure, the upper-level problem aims to determine the target objective value and ensure the feasibility of the CCs for TCL scheduling, and the lower-level problem focuses on minimizing the expectation of constraint violations while adhering to the upper bound provided by the upper-level problem. We then deploy hinge-loss approximation and a convergent block coordinate descent (BCD) method to make the bilinear low-level problem tractable. The proposed approach is evaluated using real-world data, and simulation results demonstrate a notable 4.92% reduction in costs compared to the state-of-the-art CVaR approximation method. The result also shows that through the implementation of the carbon-aware algorithm, the building is able to reduce up to 7.55% of its carbon emissions.

In Table 4.1, we present a comparison between our work and closely related studies in the literature. In the endeavor to formulate a carbon-aware optimization frame-

Table 4.1: Comparison between different approaches

Method	Carbon-aware	Distribution Assumption-Free	Low Run-time Complexity	Conservatism
SO[177] [90]	×	×	√	Low
RO[63] [73] [60]	×	√	√	Very High
CC[154]	×	×	√	Low
DRO[144]	×	√	√	High
DRCC-CVaR[71] [173]	×	√	√	Medium
DRCC-MILP[44][61]	×	√	×	Low
Ours	√	√	√	Low

work for decision-making regarding TCL control, our primary contributions can be summarized as follows:

1. We formulate a carbon-aware model for optimal TCL scheduling that explicitly accounts for the carbon emissions involved in TCL operation.

2. We develop a bilevel optimization reformulation strategy to tackle uncertainties inherited in the TCL scheduling.

3. We conduct the theoretical analysis to illustrate that the proposed approach is guaranteed to exceed the performance of the conventional CVaR approximation method.

4. We perform extensive simulation studies employing real-world datasets to shed light on how the proposed approach can contribute to building decarbonization.

The subsequent sections of this paper proceed as follows: Section 4.2 discusses the related work. In Section 4.3, we introduce the deterministic carbon-aware TCL control model. Section 4.4 outlines the conventional CVaR approximation of CCPs and provides an overview of the proposed bi-level framework. In Section 4.5, we present the DRCC approach for TCL control and its corresponding solutions. Our numerical studies are detailed in Section 4.6, and finally, Section 5.7 draws our conclusions.

4.2 Related Work

4.2.1 TCLs Control

To continuously control individual components involved in TCLs, recently a strategic control learning framework for HVAC systems using deep reinforcement learning (DRL) has been proposed in [127]. To integrate individual TCLs located in different zones of a building into building-level operations, previous research efforts, such as [90], explored minimizing the aggregate cost associated with TCLs by employing weighting factors to balance cost savings and thermal comfort. When aggregated, TCLs have the capability to participate in energy price arbitrage and provide reserve services to the power system. The operational flexibility of TCLs is explored in [88] through the modeling of a risk-averse aggregator that manages decentralized TCLs with the objective of maximizing profit. This approach employs a high-level thermal energy storage model to quantify the flexibility potential of TCL aggregations. The flexibility of aggregated TCLs is characterized in [68], where constraints on temperature, cycling, and total energy consumption are incorporated. Additionally, a study on event-triggered power tracking control for heterogeneous populations of thermostatically controlled loads (TCLs) is presented in [194].

4.2.2 Handling Uncertainties of Prediction Errors

Furthermore, forecasting errors of ambient temperature involved in TCL scheduling have been examined in the literature using both stochastic optimization (SO) [177] [90] and robust optimization (RO) methods [60]. However, SO necessitates the assumption of a specific probability distribution, while the results of RO can be too conservative. An alternative is chance-constrained programming (CCP), which guarantees the probability of occupant comfort under the worst probability distribution is still greater than the threshold [164] and strikes a balance between robustness and conservatives. However, CCP still requires the knowledge about probability distribution for the random variables involved in the TCL scheduling problem [54]. To address this limitation and alleviate the need for knowledge about the distributions, distributionally robust chance-constrained (DRCC) methods have been proposed to facilitate decision-making over a family of distributions, termed ambiguity set, that is supported by empirical data [183]. Nevertheless, addressing complex CCPs remains a challenge considering their non-differentiability and non-convexity.

In prior research, efforts have been made to solve data-driven DRCC TCL control through the use of its Conditional Value at Risk (CVaR) approximation [71] [173]. The advantage of the CVaR approximation lies in its ability to maintain linearity and incorporate supporting information, albeit potentially resulting in a conservative solution [126]. The increased conservativeness arises from the CVaR's inherent consideration of the severity of violations, leading to a lower violation probability in comparison to predefined thresholds [44]. An alternative approach is presented in [61], offering an exact deterministic reformulation for data-driven CCPs within Wasserstein balls. However, it necessitates transforming the problem into Mixed-Integer Programs (MIPs), which can be computationally prohibitive, particularly for larger problems [62].

4.3 System Model

In this section, we first present the deterministic method of carbon-aware TCL control, then proceed to discuss the joint chance-constrained programming formulation designed to handle temperature uncertainty. Figure 4.1 illustrates the optimization framework of the system.

4.3.1 Deterministic Carbon-Aware TCLs Model

1) *TCLs Model*

Consider a group of TCLs, which is composed of multiple individual load devices. Each TCL system i is fully independent of all the other TCL units. The TCL system model is given by the following equations which have been illustrated in existing paper [85]:

$$T_{i,t+1} = \alpha T_{i,t} + (1-\alpha)\left(T_t^a - \eta R q_{i,t}\right), \forall t, \tag{4.1}$$

where $T_{i,t}$ is the indoor temperature of zone i at time t. T_t^a is the outdoor temperature, and η denotes the coefficient of performance of TCL, which is positive for cooling and negative for heating. R represents its thermal resistance. The TCL parameter α is introduced to account for the influence of the current room temperature on the room temperature at the next time slot. It is defined as $e^{-\frac{\Delta t}{CR}}$, where C represents the thermal capacitance of the TCL, and Δt represents the duration of a time step. The TCL scheduling problem is formulated in a rolling procedure, where the optimization model is executed every N time slot using a moving window.

We suppose that the power consumption satisfies the upper limit and the lower limit, and we have:

$$\sum\nolimits_{i=1}^{N} q_{i,t} \leq L_r, \forall t, \tag{4.2}$$

where the threshold value L_r is imposed by the characteristics of the power distribution feeder/bus of the building [75]. This threshold is assumed to remain the same at each period t. The primary objective of TCL systems is to regulate the indoor air temperature within a specified comfort zone, which can either be a fixed range or vary over time. The constraints on the indoor air temperature $T_{i,t}$ are defined by

$$T_t^{lb} \leq T_{i,t} \leq T_t^{ub}, \forall t, \tag{4.3}$$

where T_t^{lb} and T_t^{ub} represent the lower and upper bounds of the comfort zone during the time interval t.

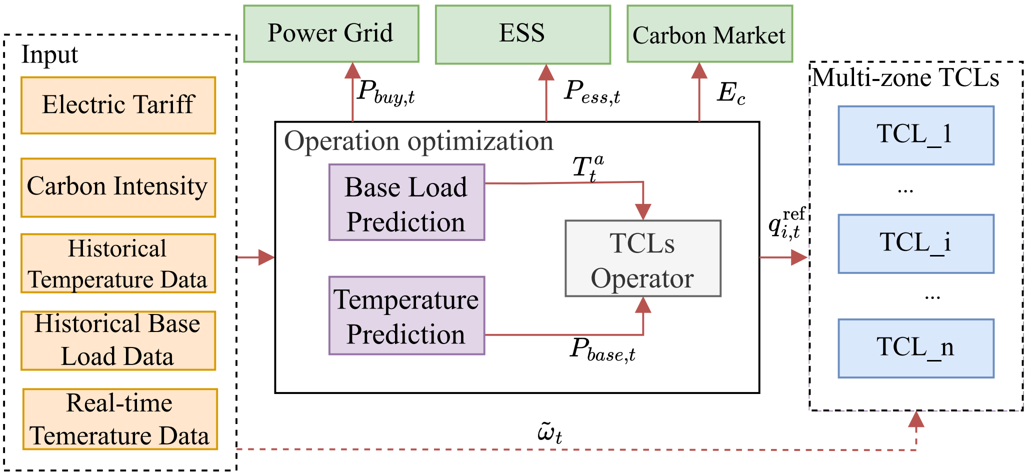


Figure 4.1: Illustration of the proposed optimization framework.

2) *Carbon Emission Model*

In this study, we consider that carbon emissions are directly linked to power consumption. Rather than depending on a constant factor for calculating carbon emissions originating from the grid, as indicated by previous studies [176], our study adopts a more accurate carbon accounting model involving the concept of carbon intensity [185]. It measures the carbon emissions produced per unit of electricity generated. Then we have,

$$\sum\nolimits_{t=1}^{T} CI_t \cdot P_{buy,t} \cdot \Delta t, \tag{4.4}$$

where $P_{buy,t}$ denotes the power of the energy purchased of the building at time slot t. CI_t is the carbon intensity of the main grid at time t, which is determined as following [121]:

$$CI_t = \frac{\sum_{k \in \mathcal{E}} ef^k \times M_t^k}{\sum_{k \in \mathcal{E}} M_t^k}, \tag{4.5}$$

where $\mathcal{E}$ denotes the set of energy sources, such as wind generation, solar power, and coal. M_t^k is the electricity output from source k at t. Carbon emissions resulting from the generation of electricity differ among the various energy sources. For each source

k, there exists a carbon emission factor, ef^k, which specifies the amount of emissions produced for every unit of electricity generated by that source.

CI_t represents the average carbon emissions per unit of electricity, changed on an hourly basis according to the proportionate contribution of each type of energy source. It offers a more precise measurement of carbon emissions. Consequently, CI_t is crucial for gauging the environmental footprint of power generation. Major corporations such as Google and Microsoft have already declared their intentions to engage in the hourly trading of credits for zero-carbon electricity [12]. Hourly varying carbon intensity is anticipated to become increasingly significant in the future. The composition of energy sources used in real-time is typically made available by Independent System Operators (ISOs) [6]. Furthermore, real-time data on carbon intensity can be obtained from third-party services like Electricity Map [11].

In practice, the carbon trading scheme has gained widespread recognition and has progressively become a mainstream approach for the global society to address climate change [37]. Many countries and regions, including Japan [43], New York [40], China [195], and Germany [53], have legislated the inclusion of specific types of buildings, such as large public buildings and commercial buildings, in their carbon trading programs. The Emissions Trading System (ETS) enables building owners to derive benefits from energy efficiency in energy-efficient buildings [150].

To achieve net-zero carbon emissions, the building needs to buy the carbon allowance from the carbon market. The purchased allowance is equal to:

$$E_c = \sum\nolimits_{t=1}^{T} P_{buy,t} \cdot CI_t - E_{bank}. \tag{4.6}$$

In this context, E_c represents the carbon allowance acquired from the market, while E_{bank} denotes the quantity of carbon allowance held by the building. The objective is to incentivize the building to decrease its energy consumption or time-shift its load to periods when carbon intensity is low, such as when the majority of energy is derived from renewable sources.

3) *Power Balance and ESS Model*

The power balance constraints are:

$$P_{buy,t} = \sum\nolimits_{i=1}^{N} q_{i,t} + P_{base,t} + P_{ess,t} \tag{4.7}$$

where $P_{base,t}$ denotes the energy consumption excepts TCLs. $P_{ess,t}$ is charging and discharging power of ESS at time t (positive when charging and negative when discharging). The energy storage system (ESS) is modeled as:

$$Q_{ess,t} = Q_{ess,t-1} + P_{ess,t}, \tag{4.8}$$

$$P_{ess,min} \leq P_{ess,t} \leq P_{ess,max}, \tag{4.9}$$

$$Q_{ess,min} \leq Q_{ess,t} \leq Q_{ess,max}, \tag{4.10}$$

where $P_{ess,min}$ and $P_{ess,max}$ are the minimum and maximum charge and discharge power of the ESS, respectively. $Q_{ess,t}$ denotes the electricity storage level of the ESS at t.

Therefore, the overall deterministic carbon-aware TCL scheduling problem is formulated as follows:

$$\begin{aligned} (DO) \qquad & \min_{q_{i,t}, P_{buy,t}, P_{ess,t}, E_c} \sum\nolimits_{t=1}^{T} c_t \cdot P_{buy,t} \cdot \Delta t + E_c \cdot p_{CO_2} \\ \text{s.t.} \qquad & (4.1) - (5.16), \end{aligned} \tag{4.11}$$

where c_t denotes the electricity price at time t. p_{CO_2} denotes the carbon price.

The deterministic model exhibits limitations in terms of its robustness, as it relies exclusively on predictions of the outdoor air temperature and base load for the next N hours to calculate the schedules. When the actual temperature and load deviate from these predictions, it undermines the feasibility and optimality of the schedule. To strengthen the robustness of optimal schedules, we introduce the prediction error of outdoor air temperature and load as stochastic variables. In the subsequent subsections, we delve into the predominant formulations employing chance-constrained programs to tackle this challenge.

4.3.2 Chance-Constrained Formulation

Due to the prediction error in scheduling, the power consumption of TCL is adjusted automatically according to the forecast error of the outdoor temperature [71] [173]:

$$q_{i,t} = q_{i,t}^{\mathrm{ref}} + \frac{1}{\eta \cdot R_i} \cdot \tilde{\omega}_t, \tag{4.12}$$

where $\tilde{\omega}_t$ denotes the prediction error of the outdoor temperature T_t^a at time t. If $\omega_t \geq 0$, indicating that the predicted T_t^a is lower than the actual value, it becomes necessary to increase the power output in order to restore the temperature to the comfortable zone, and vice versa. To eliminate the impact of outdoor temperature prediction errors at different time slots, the indoor temperature will remain the same when the energy consumption power of TCL is adjusted in real time according to the forecast error of the outdoor temperature.

In the presence of uncertainty, the base load is formulated as:

$$P_{base,t} = \overline{P}_{base,t} + \tilde{P}_{base,t}, \tag{4.13}$$

where $\overline{P}_{base,t}$ is the day-ahead predicted value of base load. The random terms are denoted by $\tilde{P}_{base,t}$. Due to the existence of uncertainty, the joint chance constraints are modeled as follows:

$$\mathbb{P}\left(\begin{array}{l} \sum_{i=1}^{N} q_{i,t} \leq L_r, \forall t \\ q_{i,t} \leq q_{i,max}, \forall i, \forall t \\ \sum_{t=1}^{T} \sum_{i=1}^{N} q_{i,t} + P_{base,t} + P_{ess,t} \leq P_{buy,t}, \forall t \end{array}\right) \geqslant 1 - \epsilon. \tag{4.14}$$

The primary objective of the power balance chance constraint in the system model is to maximize system reliability, ensuring that the load demand is met with high probability. In cases of excess power supply, a dumping load can be introduced to absorb the surplus, maintaining power balance in practical system operations [70].

Then the chance-constraint net-zero building TCL control problem is:

$$(CC) \quad \min_{q_{i,t}^{ref}, P_{buy,t}, P_{ess,t}, E_c} \sum\nolimits_{t=1}^{T} c_t \cdot P_{buy,t} \cdot \Delta t + E_c \cdot p_{CO_2}$$
$$\text{s.t. } (4.1) - (5.16), (4.12) - (4.14). \quad (4.15)$$

Given the existence of uncertainty, it becomes impractical to enforce deterministic constraints on random variables. Consequently, chance constraints are adopted to characterize the power balance and TCL energy consumption limitations. The incorporation of chance constraints presents distinct advantages compared to alternative approaches, as it circumvents the need to account for rare scenarios, ultimately yielding enhanced performance in scenarios that are commonly encountered and have a high likelihood of occurrence.

4.4 Data-Driven Chance constrained programs for TCL scheduling

To deal with the joint chance constraints, in this section, we begin by presenting the Conditional Value at Risk (CVaR) formulation of the Chance-Constrained Program (CCP). Subsequently, we introduce the proposed bi-level reformulation strategy for the CCP problem.

4.4.1 Traditional CVaR Approximation for CCP

Observing that all constraints inside the probability operator of Eq. (4.14) are linear, they can be expressed in a generic form:

$$(\boldsymbol{H}_m \tilde{\boldsymbol{\xi}})^\top \boldsymbol{x} \leq b_m(\boldsymbol{x}), \forall m \in \mathcal{M}, \quad (4.16)$$

where $\boldsymbol{H}_m$ is the coefficient matrix in each constraint. $\tilde{\boldsymbol{\xi}} = [\tilde{\boldsymbol{\omega}}_t, \tilde{\boldsymbol{P}}_{base,t}]$ denotes the random variable. $\boldsymbol{x} = \left[\boldsymbol{q}_{i,t}^{ref}, \boldsymbol{P}_{buy,t}, \boldsymbol{P}_{ess,t}, \boldsymbol{E}_c\right]$ represents the decision variable vector. b_m is an affine function of $\boldsymbol{x}$. And $m \in \mathcal{M} = \{1, 2, \cdots, M\}$ is the index of the constraints inside the probability operator. We define a new function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ as

$$h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) = \max_{m \in \mathcal{M}}((\boldsymbol{H}_m \tilde{\boldsymbol{\xi}})^\top \boldsymbol{x} - b_m(\boldsymbol{x})). \tag{4.17}$$

The m-th constraint $h_m(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) : \mathbb{R}^n \times \Xi \to \mathbb{R}$ for all $m \in \mathcal{M}$ and $h_m(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is convex and lower semicontinuous in $\boldsymbol{x}$ for every $\tilde{\boldsymbol{\xi}} \in \Xi$.

Therefore, Eq. (4.14) can be expressed as

$$\mathbb{P}(h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0) \geq 1 - \epsilon \tag{4.18}$$

For convenience, the joint chance constraint program problem (4.15) can be written as a standard CCP:

$$Z^* = \min_{x \in \mathcal{X}} \left\{ f(\boldsymbol{x}) : \mathbb{P}\{\tilde{\boldsymbol{\xi}} : h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0\} \geq 1 - \varepsilon \right\}, \tag{4.19}$$

where $\boldsymbol{x}$ in $\mathbb{R}^n$ is the decision variables. The objective function $f(\boldsymbol{x})$ is convex. $\tilde{\boldsymbol{\xi}}$ is the random vectors. Function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) = \max_{i \in [I]} h_i \ (\boldsymbol{x}, \boldsymbol{\xi})$, where $h_i(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) : \mathbb{R}^n \times \Xi \to \mathbb{R}$ for all $i \in [I] := \{1, \ldots, I\}$ and $h_i(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is convex and lower semicontinuous in $\boldsymbol{x}$ for almost every $\tilde{\boldsymbol{\xi}} \in \Xi$. Given the random variable $\tilde{v}$ and a risk level ε, the conditional value at risk (CVaR) is defined as:

$$\mathrm{CVaR}_{1-\varepsilon}(\tilde{v}) := \min_{\beta} \left\{ \beta + \frac{1}{\varepsilon} \mathbb{E}_{\mathbb{P}}[\tilde{v} - \beta]_+ \right\}. \tag{4.20}$$

The CVaR approximation of the CCP (4.19) is formulated as following:

$$Z^{\mathrm{CVaR}} = \min_{x \in \mathcal{X}} \left\{ f(\boldsymbol{x}) : \min_{\beta \leq 0} \left\{ \beta + \frac{1}{\varepsilon} \mathbb{E}\{h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta\}_+ \right\} \leq 0 \right\}. \tag{4.21}$$

The CVaR approximation-based chance constraints TCL control has been well studied [71] [173]. However, it usually returns a suboptimal solution. In the next subsection, we introduce two new convex approximations for the chance-constrained TCL scheduling problem.

4.4.2 Proposed Bilevel Optimization Framework for CCP

Consider the CCP (4.19), in fact, $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0\}$ is equal to $\mathbb{E}_{\mathbb{P}}[\mathbb{I}(h(x, \tilde{\boldsymbol{\xi}}) \leq 0)]$, where $\mathbb{I}(\cdot)$ is the indicator function. Thereby we have the following equivalent formulation:

Theorem 4.4.1 (Equivalent formulation of CCP [94]). *The CCP (4.19) can be viewed as following equivalent formulation*

$$Z^* = \min_{x \in \mathcal{X}, \tau} \Big\{ \tau : h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \geq 1 - \varepsilon,$$
$$\mathbb{E}[z(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}})] = 0, z(\tilde{\boldsymbol{\xi}}) \in [0, 1], s(\tilde{\boldsymbol{\xi}}) \geq 0, f(\boldsymbol{x}) \leq \tau \Big\}. \tag{4.22}$$

where τ is a auxiliary variable which replacing the objective function, $z(\cdot) : \Xi \to \mathbb{R}$ is a functional variable representing the indicator function $\mathbb{I}(\cdot)$. $s(\cdot) : \Xi \to \mathbb{R}$ is an auxiliary nonnegative functional variable denoting uncertain constraint violations.

Note that $\mathbb{I}(g(x, \tilde{\boldsymbol{\xi}}) \leq 0) \geq z(\tilde{\boldsymbol{\xi}})$ is replaced by $\mathbb{E}[z(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}})] = 0$. And the functional variable $z(\cdot)$ can be either binary or continuous. More details of this proof can be found in [94].

black Problem (4.22) suggests that for a CCP, it is possible to search for the smallest possible objective value while maintaining the feasibility of the constraints. This encourages us to transform the CCP into a bi-level optimization problem in the form of:

$$Z^* = \min_{\tau} \tau, \tag{4.23}$$
$$\text{s.t.} \quad (x^*, s^*(\cdot), z^*(\cdot)) \in \arg\min_{\substack{x \in \mathcal{X} \\ z(\cdot) \in [0,1] \\ s(\cdot) \geq 0}} \Big\{ \mathbb{E}[z(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}})] :$$
$$h(x, \tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \geq 1 - \varepsilon, f(\boldsymbol{x}) \leq \tau \Big\}, \tag{4.24}$$
$$\mathbb{P}\Big\{ \tilde{\boldsymbol{\xi}} : h\Big(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}\Big) \leq 0 \Big\} \geq 1 - \varepsilon, \tag{4.25}$$

where Eq. (4.23) and Eq. (4.25) are identified as an upper-level problem with the decision variable τ. Eq. (4.24) can be considered as the lower-level problem with decision variables $x, s(\cdot)$, and $z(\cdot)$. The formulation in Eq. (4.22) and the bi-level optimization formulation in Eqs. (4.23) - (4.25) are equivalent, according to the theoretical proof provided in [94]. The upper-level problem aims to determine the optimal objective function value while ensuring the feasibility of a chance-constrained program for a given decision derived from the lower-level problem. The lower-level problem focuses on minimizing the expectation of constraint violations while adhering to the upper bound on the objective function value provided by the upper-level problem.

The key idea of this formulation is that given the value of the upper-level decision τ, we can solve the lower-level problem in Eq. (4.24) and then verify if the solution meets the condition Eq. (4.25). If the answer is affirmative, we can decrease the value of τ; if not, we increase τ. The optimal upper-level decision τ^* can be determined using a binary search. Then the challenge lies in solving the lower-level problem for a given objective target τ since it is a two-stage bilinear program, which is known as difficult to solve. To mitigate this issue, we substitute the bilinear program Eq. (4.24) with a hinge-loss approximation. Specifically, the lower-level problem in Eq. (4.24) can be simplified by setting the functional variable $z(\boldsymbol{\xi}) = 1$ [36], which reduces the problem to the following form:

$$Z^A := \min_{x \in \mathcal{X}, s(\cdot) \geq 0} \left\{ \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] : h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), f(\boldsymbol{x}) \leq \tau \right\}. \tag{4.26}$$

The next step is to eliminate the slack variable $s(\cdot)$ by projecting it out, leading to the following hinge-loss approximation:

$$Z^A = \min_{x \in \mathcal{X}} \left\{ \mathbb{E}\left[h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})_+ \right] : f(\boldsymbol{x}) \leq \tau \right\}. \tag{4.27}$$

where where $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})_+$represents the nonnegative part of the function $h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$, defined as:

$$h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})_+ = \max\{0, h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}. \tag{4.28}$$

The objective function in Eq. (4.27), referred to as the hinge-loss function, can be interpreted as an expectation of the nonnegative part of the random function $h(x, \tilde{\boldsymbol{\xi}})$. The aim of the hinge-loss approximation is to minimize the expectation of infeasibilities, given the decision τ of the upper-level problem. By substituting the lower-level problem in Eq. (4.24) with the hinge-loss approximation in Eq. (4.27), we now have the following problem:

$$
\begin{aligned}
Z^A = & \min_{\tau} \tau, && (4.29)\\
& \text{s.t.} \quad (x^*, s^*(\cdot)) \in \underset{x\in\mathcal{X}, s(\cdot)\geq 0}{\arg\min} \Big\{ \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] : h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), \\
& \qquad\qquad f(\boldsymbol{x}) \leq \tau \Big\}, && (4.30)\\
& \mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} : s^*(\tilde{\boldsymbol{\xi}}) = 0 \right\} \geq 1 - \varepsilon. && (4.31)
\end{aligned}
$$

If the optimal solution x^* of problem (4.29) - (4.31) is feasible of to CCP (4.19), then we have $\mathbb{P}\{\tilde{\xi} : h(x^*, \tilde{\boldsymbol{\xi}}) \leq 0\} \geq 1 - \varepsilon$, which is equivalent to $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : s^*(\tilde{\boldsymbol{\xi}}) = 0\} \geq 1 - \varepsilon$.

In the data-driven scenario, the probability distribution of the random variable is assumed to have finite support with N scenarios. Specifically, here we assume that the random variable $\tilde{\boldsymbol{\xi}}$ has N historical sample $\Xi = \left\{ \boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^N \right\}$ with $\mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^i \right\} = 1/N$ for all $i \in [N]$. Under this assumption, the chance constraint (4.19) reduces to:

$$
\sum_{i\in[N]} \mathbb{I}\left(h\left(\boldsymbol{x}, \boldsymbol{\xi}^i \right) \leq 0 \right) \geq N - \lfloor N\varepsilon \rceil. \tag{4.32}
$$

The new problem can be expressed in the following form:

$$
\begin{aligned}
Z^A = & \min_{\tau} \quad \tau && (4.33)\\
& \text{s.t.} \quad (x^*, s^*) \in \underset{x\in\mathcal{X}, s\geq 0}{\arg\min} \Big\{ \frac{1}{N} \sum_{i\in[N]} s_i : f(\boldsymbol{x}) \leq \tau, \\
& \qquad\qquad h\left(\boldsymbol{x}, \boldsymbol{\xi}^i \right) \leq s_i, \forall i \in [N] \Big\}, && (4.34)\\
& \sum_{i\in[N]} \mathbb{I}\left(s_i^* = 0 \right) \geq N - \lfloor N\varepsilon \rfloor. && (4.35)
\end{aligned}
$$

Note that in the discrete support, the lower-level problem can be computed effectively. While this assumption may appear restrictive, if the possible values for ξ are generated via Monte Carlo sampling from a general distribution, the resulting problem can be viewed as an approximation of one with a general distribution [117]. There is both theoretical and empirical evidence supporting the idea that such a sample-based approximation can effectively solve problems with continuous distributions with reasonable computational effort [118]. There are also some related results on sampling approaches to chance-constrained mathematical programs [55] [56].

Binary search is then employed to identify the optimal τ. Another point is the determination of a more suitable upper bound τ^u and lower bound τ^l, as this could enhance the efficiency of the algorithms by decreasing the number of iterations. Here we choose the result of CVaR as its upper bound and a quantile bound proposed in [36] as its lower bound. When the difference between the upper bound τ^u and lower bound τ^l exceeds the threshold, we set τ to their median and obtain an optimal solution (x^*, s^*). We then verify its feasibility with respect to the chance constraints. If $\sum_{i\in\mathcal{N}} \mathbb{I}\left(s_i^* = 0\right) \geq N - \lfloor N\varepsilon \rfloor$, then x^* is feasible, and therefore τ is considered a valid upper bound. Consequently, τ^u is updated to τ, and the process is repeated to seek a better feasible solution, thereby improving the upper bound. Conversely, if $\sum_{i\in\mathcal{N}} \mathbb{I}(s_i^* = 0) \leq N - \lfloor N\varepsilon \rfloor$, τ^l is updated to τ, and the steps are repeated in an attempt to find a feasible solution.

black Overall, the proposed bi-level building TCL control algorithm can be described in Algorithm 3.

Theorem 4.4.2 (Performance guarantee [94]). *Let Z^{CVaR} denote the optimal value of the CVaR approximation. When Set $\mathcal{X}$ is nonempty, closed, convex, contained in a closed convex cone, we must have $Z^A \leq Z^{\mathrm{CVaR}}$.*

Proof. The proof of this theorem can be found in [94]. □

Algorithm 3 Proposed Hinge-loss (HL)-based Bi-level Reformulation of the TCL Scheduling Algorithm

Input: Stopping tolerance δ, base load prediction error samples $[\widetilde{P}_{base,t,1}, ..., \widetilde{P}_{base,t,N}]$, temperature prediction samples$[\tilde{w}_{t,1}, ..., \tilde{w}_{t,N}]$, violation level ε;

Output: Optimal solution $\left\{q^*_{i,t}, P^*_{buy,t}, P^*_{ess,t}, E^*_c, s^*(\cdot)\right\}$ and its optimal objective value Z^A;

Solve the CVAR-based TCL scheduling problem (4.21), get the objective value Z^{CVaR}, and set it the upper bound of the problem $\tau^u \leftarrow Z^{CVaR}$;

Solve the problem for each sample $[\widetilde{P}_{base,t,i}, \tilde{w}_{t,i}]$, $i \in N$ and get the objective value Z^i,then sort the result and get the set $[Z^1, ..., Z^N]$ in ascending order. Then set the lower bound $\tau^l \leftarrow Z^k$, where $k = N - \lfloor N\varepsilon \rfloor$;

while $\tau^u - \tau^l > \delta$ **do**

    Set $\tau \leftarrow (\tau^l + \tau^u)/2$, and calculate get the optimal solution $\left\{q^*_{i,t}, P^*_{buy,t}, P^*_{ess,t}, E^*_c, s^*(\cdot)\right\}$ of the hinge-loss approximation (4.27);

    if $\sum_{i\in[N]} \mathbb{I}\left(s^*_i = 0\right) \geq N - \lfloor N\varepsilon \rfloor$ **then**

        $\tau^l \leftarrow \tau$;

    else

        $\tau^u \leftarrow \tau$;

    end if

end while

Return the optimal solution $\left\{q^*_{i,t}, P^*_{buy,t}, P^*_{ess,t}, E^*_c, s^*(\cdot)\right\}$ and its optimal objective value Z^A.

CVaR approximation is well-known as the best convex approximation of CCP, but the solution is usually sub-optimal. Theorem 2 gives the theoretical guarantee that for the TCL scheduling problem here, the HL-based methods always outperform the CVaR approximation. This is mainly due to the fact that, for an identical upper-level target τ, the feasibility check condition of the HL method is less stringent. Although the binary search of the HL approximation will slightly increase the running time, the improvement in solution quality makes the additional computation cost worthwhile.

4.4.3 Block Coordinate Descent (BCD)-based Approximation

Although the original hinge-loss approximation (4.27) which sets the functional variable $z(\boldsymbol{\xi}) = 1$ effectively approximates the original CCP problem, under certain circumstances, it may fail to discover any feasible solution. Therefore, this subsection discusses a modified strategy for Algorithm 3, employing the block coordinate descent (BCD) method to optimize $z(\cdot)$ as well in cases where no feasible point exists for Algorithm 3.In the literature, the difference-of-convex (DC) [159] approach is employed to address the lower chance-constraints problem by expressing the non-convex term as the difference between two convex quadratic functions. However, as demonstrated in [94], the BCD method can find a better solution compared to the DC method in this context.

The constraints in problem (4.24) can be divided into two parts, the functional variable $z(\cdot)$ and variables x and $s(\cdot)$. This division enables us to iteratively optimize over $z(\cdot)$ and $(x, s(\cdot))$. Specifically, at iteration $k+1$, we initially set $(x, s(\cdot)) = \left(x^k, s^k(\cdot)\right)$, where $\left(x^k, s^k(\cdot)\right)$ represents the solutions obtained from the previous iteration, in the context of the lower-level problem. We then proceed to solve the following problem:

$$z^{k+1}(\cdot) \in \arg\min \left\{ \mathbb{E}\left[z(\tilde{\boldsymbol{\xi}}) s^k(\tilde{\boldsymbol{\xi}}) \right] : z(\tilde{\boldsymbol{\xi}}) \in [0,1], \right.$$
$$\left. \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \geq 1 - \varepsilon \right\}, \tag{4.36}$$

It is achieved by sorting the values of $\left\{ s^k(\boldsymbol{\xi}) \right\}_{\xi \in \Xi}$. Subsequently, by fixing the value of the functional variable $z(\cdot) = z^{k+1}(\cdot)$, we address the subsequent optimization problem:

$$\left(x^{k+1}, s^{k+1}(\cdot) \right) \in \arg\min_{x \in \mathcal{X}, s(\cdot)} \left\{ \mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}}) \right] : f(\boldsymbol{x}) \leq \tau, \right.$$
$$h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), s(\tilde{\boldsymbol{\xi}}) \geq 0\} \tag{4.37}$$

The algorithm is detailed in Algorithm 4, and its convergence properties are discussed in [94]. And the hybrid BCD-based building TCL control algorithm is described in Algorithm 5.

Algorithm 4 Block Coordinate Descent (BCD) Method for (4.24)

Input: Stopping tolerance σ, initial result τ^0, z^0, s^0, x^0

1: **while** $\left| \mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}}) s^{k+1}(\tilde{\boldsymbol{\xi}}) \right] - \mathbb{E}\left[z^k(\tilde{\boldsymbol{\xi}}) s^k(\tilde{\boldsymbol{\xi}}) \right] \right| \geq \sigma$ **do**

2: Fix $(x, s(\cdot)) = \left(x^k, s^k(\cdot) \right)$, solve problem (4.36) and get $z^{k+1}(\cdot)$

3: Fix $z(\cdot) = z^{k+1}(\cdot)$, solve problem (4.37) and get $\left(x^{k+1}, s^{k+1}(\cdot) \right)$

4: **end while**

5: Return the result $(x^{k+1}, z^{k+1}(\cdot), s^{k+1}(\cdot))$

4.5 Data-Driven Distributionally Robust Chance Constrained Approach for TCL Scheduling

In practice, the complete distributional information of random parameters may not be fully known. To address this challenge and hedge against the uncertainty of dis-

Algorithm 5 BCD-based TCL Scheduling Algorithm

Input: Stopping tolerance δ, base load prediction error samples $[\widetilde{P}_{base,t,1}, ..., \widetilde{P}_{base,t,N}]$, temperature prediction samples$[\tilde{w}_{t,1}, ..., \tilde{w}_{t,N}]$, violation level ε;

Output: Optimal solution $\left\{q^*_{i,t}, P^*_{buy,t}, P^*_{ess,t}, E^*_c, s^*(\cdot)\right\}$ and its optimal objective value Z^{A+};

1: Solve the CVaR-based TCL scheduling problem (4.21), get the objective value Z^{CVaR}, and set it the upper bound of the problem $\tau^u \leftarrow Z^{CVaR}$;

2: Solve the problem for each sample $[\widetilde{P}_{base,t,i}, \tilde{w}_{t,i}]$, $i \in N$ and get the objective value Z^i, then sort the result and get the set $[Z^1, ..., Z^N]$ in ascending order. Set the lower bound $\tau^l \leftarrow Z^k$ where $k = N - \lfloor N\varepsilon \rfloor$;

3: **while** $\tau^u - \tau^l > \delta$ **do**

4:     Set $\tau \leftarrow (\tau^l + \tau^u)/2$, and calculate get the optimal solution $\left\{q^*_{i,t}, P^*_{buy,t}, P^*_{ess,t}, E^*_c, s^*(\cdot)\right\}$ of the hinge-loss approximation (4.27);

5:     **if** $\sum_{i\in[N]} \mathbb{I}\left(s^*_i > 0\right) \le N - \lfloor N\varepsilon \rfloor$ **then**

6:         $\tau^u \leftarrow \tau$

7:     **else**

8:         Solve the alternating minimization method algorithm 4 for current $\left\{q^*_{i,t}, P^*_{buy,t}, P^*_{ess,t}, E^*_c, s^*(\cdot)\right\}$, and get new result $\left\{q^{*+}_{i,t}, P^{*+}_{buy,t}, P^{*+}_{ess,t}, E^{*+}_c, s^{*+}(\cdot)\right\}$

9:         **if** $\sum_{i\in[N]} \mathbb{I}\left(s^*_i > 0\right) \le N - \lfloor N\varepsilon \rfloor$ **then**

10:             $\tau^u \leftarrow \tau$;

11:         **else**

12:             $\tau^l \leftarrow \tau$;

13:         **end if**

14:     **end if**

15: **end while**

16: Return the optimal solution $\left\{q^*_{i,t}, P^*_{buy,t}, P^*_{ess,t}, E^*_c, s^*(\cdot)\right\}$ and its optimal objective value Z^{A+}

tributional information, we explore the concept of distributionally robust chance-constrained (DRCC) problems. They ensure that the chance constraint is satisfied across all probability distributions within a specified family, known as the ambiguity set. This section begins with an introduction to DRCC approaches, followed by an exposition of its CVaR approximation. Additionally, we discuss the bi-level approximation for the DRCC-TCL scheduling problem.

4.5.1 DRCC with Wasserstein Distance

Consider a standard DRCC optimization problem:

$$Z^* = \min_{x \in \mathcal{X}} \left\{ f(\boldsymbol{x}) : \inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}\{\tilde{\boldsymbol{\xi}} : h(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0\} \geq 1 - \varepsilon \right\} \tag{4.38}$$

The random vectors $\tilde{\boldsymbol{\xi}}$ is supported on set Ξ with a joint probability distribution $\mathbb{P}$ from a family $\mathcal{P}$, termed ambiguity set. The choice of ambiguity sets $\mathbb{P}$ varies in different applications. In data-driven optimization, a popular ambiguity set is constructed using the Wasserstein metric, defined as follows:

Definition 4.5.1 (The q-Wasserstein ambiguity set)**.**

$$\mathcal{P}_q^W = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathcal{Z}) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{\xi}} \sim \mathbb{P} \\ \Delta(\mathbb{P}, \hat{\mathbb{P}}) \leq r \end{array} \right\},$$

where $r \geq 0$ donates the Wasserstein radius. For $q \in [1, \infty]$ the q-Wasserstein distance is defined as

Definition 4.5.2 (The q-Wasserstein distance)**.**

$$\Delta(\mathbb{P}, \hat{\mathbb{P}}) = \left(\inf_{\gamma \in \Pi(\mathbb{P}, \hat{\mathbb{P}})} \int \|w - v\|^p d\gamma(w, v) \right)^{\frac{1}{q}} \tag{4.39}$$

where $\Pi(\mathbb{P}, \hat{\mathbb{P}})$ represents all the joint distributions γ for $(\mathbb{P}, \hat{\mathbb{P}})$. Specifically, in this paper, we study the ∞-Wasserstein ambiguity set. And under this setting, DRCC form (4.38) admits a neat equivalent representation.

$$Z^*_\infty = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ f(\boldsymbol{x}) : \mathbb{P}_{\tilde{\boldsymbol{\zeta}}} \left\{ \tilde{\boldsymbol{\zeta}} : \bar{h}_i(\boldsymbol{x}, \tilde{\boldsymbol{\zeta}}) \leq 0, \forall i \in [I] \right\} \geq 1 - \varepsilon \right\} \tag{4.40}$$

where the convex and lower semicontinuous function $\bar{h}_i : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is defined as $\bar{h}_i(\boldsymbol{x}, \boldsymbol{\zeta}) = \max_{\boldsymbol{\xi}} \{h_i(\boldsymbol{x}, \boldsymbol{\xi}) : \|\boldsymbol{\xi} - \boldsymbol{\zeta}\| \leq \theta\}$ for each $i \in [I]$.

Furthermore, with certain additional assumptions on the function $h_i(\cdot,\cdot)$, their robust counterparts $\bar{h}_i(\cdot,\cdot)$ can be represented in simplified forms [94]. Specifically, in the context of the TCL control problem, the norm is chosen as L_∞ (i.e., $\|\cdot\| = \|\cdot\|_\infty$) and the function $h_i(\boldsymbol{x}, \boldsymbol{\xi})$ is monotone nondecreasing in $\boldsymbol{\xi}$ for any $\boldsymbol{x} \in \mathcal{X}$ and $i \in [I]$, we have:

$$\bar{h}_i(\boldsymbol{x}, \boldsymbol{\zeta}) = h_i(\boldsymbol{x}, \boldsymbol{\zeta} + r\boldsymbol{e}), \forall i \in [I] \tag{4.41}$$

4.5.2 Traditional CVaR-based DRCC Optimization

The CVaR approximation of the DRCC problem (4.38) is defined as

$$\begin{aligned} Z^{\text{CVaR}}_\infty = & \min_{x\in\mathcal{X}} \Big\{ f(\boldsymbol{x}) : \\ & \sup_{\mathbb{P}\in\mathcal{P}^W_\infty} \inf_\beta \left[\beta + \frac{1}{\varepsilon} \mathbb{E}_{\mathbb{P}} \left[\max_{i\in[I]} \Big(h_i(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta \Big)_+ \right] \right] \leq 0 \Big\}. \end{aligned} \tag{4.42}$$

Under the $\infty-$Wasserstein ambiguity set, the CVaR approximation (4.42), which is termed CVaR-based DRCCP, is equivalent to

$$\begin{aligned} Z^{\text{CVaR}}_\infty = & \min_{x\in\mathcal{X}} \Big\{ f(x) : \\ & \min_\beta \left[\beta + \frac{1}{\varepsilon} \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left\{ \max_{i\in[I]} \left\{ \bar{h}_i(\boldsymbol{x}, \tilde{\boldsymbol{\zeta}}) \right\} - \beta \right\}_+ \right] \leq 0 \Big\}. \end{aligned} \tag{4.43}$$

4.5.3 Bi-level Reformulation of DRCC Optimization Problem

The bi-level reformulation of the DRCC problem (4.38) can be given as:

$$\begin{aligned}
&Z_{\infty}^{A} = \min_{\tau} \tau, \\
&\text{s.t.} \quad \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_{\infty}^{\mathrm{W}}} \left\{ \mathbb{E}_{\mathbb{P}} \left[\max_{i\in[I]} h_i(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})_+ \right] : f(\boldsymbol{x}) \le \tau \right\}, \\
&\inf_{\mathbb{P}\in\mathcal{P}_{\infty}^{W}} \mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} : h_i\left(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}\right) \le 0, \forall i \in [I] \right\} \ge 1-\varepsilon
\end{aligned} \tag{4.44}$$

Problem (4.44) under the $\infty-$Wasserstein ambiguity set, which is termed Hinge-loss (HL)-based DRCC, is equivalent to

$$\begin{aligned}
&Z_{\infty}^{A} = \min_{\tau} \tau, \\
&\text{s.t. } x^* \in \arg\min_{x\in\mathcal{X}} \left\{ \mathbb{E}_{\mathbb{P}} \left[\max_{i\in[I]} \bar{h}_i(x, \tilde{\boldsymbol{\xi}})_+ \right] : f(\boldsymbol{x}) \le \tau \right\}, \\
&\mathbb{P}_{\tilde{\zeta}}\left\{ \tilde{\boldsymbol{\zeta}} : \bar{h}_i\left(\boldsymbol{x}^*, \tilde{\boldsymbol{\zeta}}\right) \le 0, \forall i \in [I] \right\} \ge 1-\varepsilon.
\end{aligned} \tag{4.45}$$

It is noteworthy that both the CVaR approximation for the DRCC problem (4.43) and the HL-based DRCC problem (4.45) under the $\infty-$Wasserstein ambiguity set can be regarded as extending the application of CVaR and HL approximation to the CCP (4.40), respectively. Then the solution processes for HL-DRCC and BCD-DRCC follow the same procedures outlined in Algorithm 1 and Algorithm 3, respectively. Consequently, the findings discussed in the preceding sections are applicable to the DRCC problem presented in (4.38). Thereby the HL-based DRCC approach demonstrates superior performance over the CVaR-based DRCC under the $\infty-$Wasserstein ambiguity set.

Consequently, the overall process of the problem, including the relationship between different reformulations and their corresponding solutions, is summarized in Fig. 4.2.

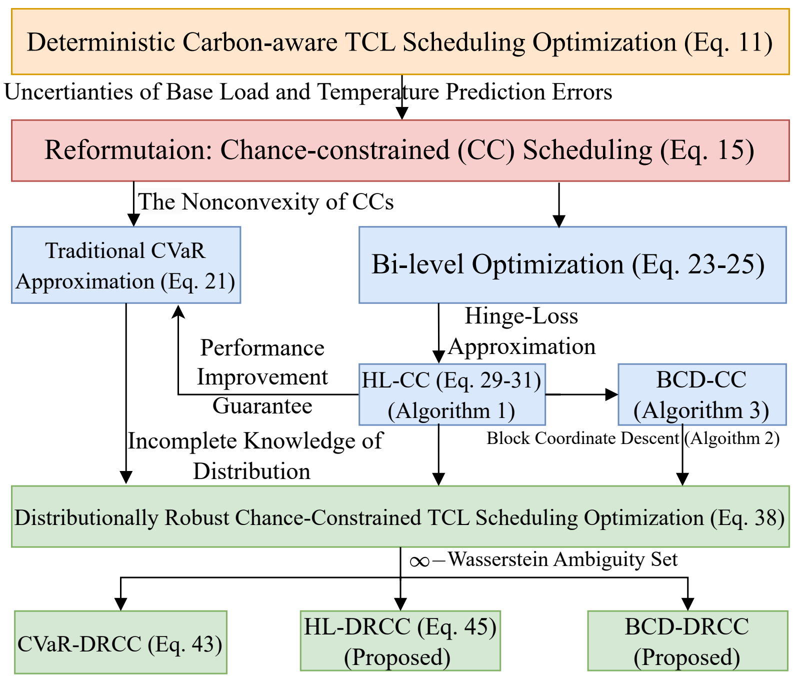


Figure 4.2: The overall process

4.6 Numerical studies

4.6.1 Data Description and Experimental Settings

In this section, we conduct comprehensive numerical studies to validate the effectiveness of the proposed algorithms. We suppose that there are 4 different zones in the building, and each has an HVAC system for cooling. The relevant parameters in the building are shown in Table 5.1. The parameters required for the HVACs are listed in Table 4.3. We set Δt to be 1 hour, and the duration of the control window N to be 24 (i.e., day-ahead scheduling). The temperature zone T_t^{lb} and T_t^{ub} are setting as 20 °C and 24 °C, respectively. The efficiency rate is set as 2.2. The $q_{i,max}$ of each TCL is set to 10 kWh. The simulations were executed on a desktop computer configured with an Intel i9 2.4GHz CPU and 32GB of RAM. The optimization problem was implemented in Python and solved using Gurobi 11.0.0.

The building load data is from a retail strip mall in Bakersfield, California on the End-Use Load Profiles for the U.S. Building Stock [161], a widely used public dataset. The temperature is from the corresponding weather data provided by the dataset. The electricity price data is sourced from [27], ensuring that our simulations are based on realistic market conditions. The carbon intensity data is obtained from Electricity Map [11], which provides real-time and day-ahead access to carbon intensity information through cloud interfaces. In Figure 4.3, the red line represents the hourly carbon intensity, while the sky blue lines represent the electricity price during the day. The base load, outdoor temperature, and carbon intensity data in different regions are provided in [1].

For synthetic scenarios, we generate the samples randomly from a multivariate Gaussian distribution with a positive-definite covariance matrix, which is widely adopted to model the load and temperature prediction error in literature [80] [95]. To test the out-of-sample performance, we generate 5000 samples. Note that in the real-world data experiments, we do not assume the distribution and use the real prediction error. And we ignore the impact of forecast error on the power limit of individual TCLs evaluating the effect of increasing TCLs. Instead, we focus on the aggregate effects on total power consumption and power balance constraints.

For real-world data experiments, to use real prediction data, we construct neural network models to forecast the day-ahead load and temperature. Because our focus is on dealing with uncertainty, we simply utilize the data from the previous day as the input of the neural network to make day-ahead predictions. By utilizing realistic data and settings, we aim to ensure the validity and practicality of our simulations, enabling meaningful insights into the performance and sensitivity of the proposed TCL scheduling model and algorithms.

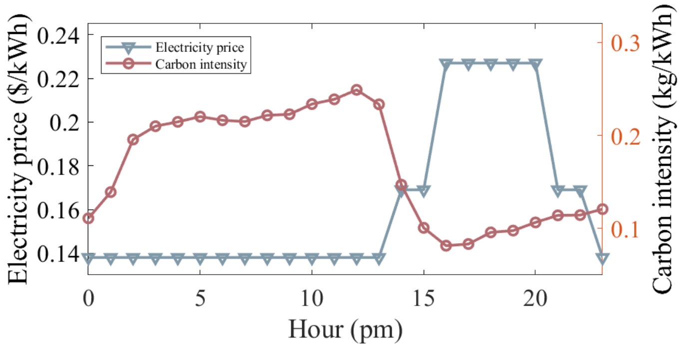


Figure 4.3: The electricity price and carbon intensity

Table 4.2: Relevant parameters in the building

Parameters	Values	Unit
T_t^{lb}	20	°C
T_t^{ub}	24	°C
η	2.2	-
$Q_{ess,max}$	100	kWh
$Q_{ess,min}$	20	kWh
$P_{ess,max}$	20	kWh
L_r	30	kWh

4.6.2 Performance Analysis for Synthetic Data

1) Run-time Analysis:

First, we evaluate the impact of the proposed method on computation time across different sample sizes. We conduct a comparison of computational efficiency among various data-driven DRCC approaches from the literature. The results in Table 4.4 indicate that the CVaR approximation-based data-driven DRCC achieves the shortest runtime. Additionally, the runtimes for the proposed HL-DRCC and BCD-DRCC methods exhibit linear growth as the sample size increases, but remain within acceptable limits. The runtime of the HL-DRCC and BCD-DRCC methods is longer than that of the CVaR approximation, which is reasonable, as they utilize the CVaR

Table 4.3: Relevant parameters for HVACs in different zones

Zone	R (°C/kW)	C (kWh/°C)	α
1	2.5	2.4	0.8464
2	1.6	2.3	0.7621
3	1.7	1.8	0.7212
4	2.4	1.5	0.7575

Table 4.4: Running time under different sample sizes

N	CVaR	HL-DRCC	BCD-DRCC	MILP-DRCC	MILP(TS)-JCC
100	0.35	0.81	0.82	27.22	18.165
200	0.73	1.64	1.71	827.86	150.53
300	0.80	2.29	2.29	6796.90	388.49
400	1.01	3.08	3.11	-	669.68
500	1.31	3.84	3.87	-	1051.87
1000	2.65	7.80	7.84	-	8942.51

approximation results as an upper bound to search for better outcomes. Although the runtime increases to some degree, the proposed method enhances the performance of the CVaR approximation, making the additional computational cost worthwhile.

Additionally, the computational efficiency of the MILP-based DRCC [44], is examined. Consistent with our prior assertion, the findings clearly show that the runtime of MILP-based DRCC becomes prohibitive for large sample sizes, attributable to the fact that solving MILP is NP-Hard. The tightening and screening (TS) method proposed in [133] aims to select appropriate Big-M values and reduce the computational burden of chance-constrained programs by eliminating superfluous constraints. While this approach has been shown to reduce the solution time of MILPs, it requires running parallel algorithms to identify the Big-M values. However, the degree of parallelism is constrained by the number of available CPU cores.

Regarding the scalability of the proposed algorithms, we evaluate the computational

Table 4.5: Running time under different Num. of TCLs

Num. of TCLs	CVaR	HL-DRCC		BCD-DRCC	
	Time	Time	Improvement	Time	Improvement
10	0.84 sec	2.36 sec	2.29%	2.38 sec	2.89%
20	1.31 sec	3.72 sec	2.25%	3.67 sec	2.55%
50	2.77 sec	7.96 sec	2.22%	8.24 sec	2.48%
100	5.21 sec	16.03 sec	2.22%	15.85 sec	2.61%

performance of the proposed approach with varying numbers of TCLs, using a fixed sample size of 500. The results in Table 4.5 demonstrate that the runtimes for the proposed HL-DRCC and BCD-DRCC methods increase linearly as the number of TCLs grows. Despite this increase, the runtimes remain within acceptable limits, indicating that the proposed approach is computationally feasible for large-scale problems.

2) Out-of-Sample Performance:

Then we evaluate the performance of the impacts of the sample sizes and Wasserstein radius on out-sample performance. We denote $\hat{\epsilon}$ as the estimated expected violation probability of the obtained solution. Table 4.6 presents the impact of sample size under an uncertainty set radius of $10\%\xi$ and a violation probability ϵ of 0.1. The results show that as the sample size increases, the out-of-sample performance improves, reflected by a reduction in violation rates due to the model gaining more information about the distribution.

Additionally, Table 4.7 demonstrates that as the uncertainty set radius increases, the objective function value increases, while the violation rate decreases, as the larger radius leads to a more conservative uncertainty set. These results highlight the trade-off between sample size, uncertainty set radius, and out-of-sample performance. The robustness of the algorithm can be adjusted by tuning both the sample size and the radius, balancing performance and conservatism.

Table 4.6: Out-of-sample performance under different sample sizes

N	CVaR		HL-DRCC		BCD-DRCC	
	Obj.	$\hat{\epsilon}$	Obj.	$\hat{\epsilon}$	Obj.	$\hat{\epsilon}$
100	\$101.64	0.0876	\$100.86	0.1109	\$99.22	0.1505
200	\$104.42	0.0298	\$102.03	0.0596	\$101.52	0.0668
300	\$104.37	0.0230	\$101.91	0.0582	\$101.51	0.0628
400	\$104.72	0.0218	\$102.25	0.0492	\$102.03	0.0540
500	\$104.71	0.0224	\$102.36	0.0464	\$102.15	0.0502
1000	\$105.24	0.0176	\$102.85	0.0390	\$102.58	0.0423

Table 4.7: Out-of-sample performance under different radius r

r	CVaR		HL-DRCC		BCD-DRCC	
	Obj.	$\hat{\epsilon}$	Obj.	$\hat{\epsilon}$	Obj.	$\hat{\epsilon}$
0.5%	\$100.48	0.0891	\$98.13	0.1802	\$97.92	0.1932
1%	\$100.70	0.0820	\$98.35	0.1696	\$98.14	0.1804
5%	\$102.48	0.0476	\$100.13	0.0958	\$99.93	0.1016
10%	\$104.71	0.0224	\$102.36	0.0464	\$102.15	0.0502
20%	\$109.17	0.0042	\$106.82	0.0081	\$106.61	0.0088

4.6.3 Performance Analysis for Real Data

1) *Comparing with performance benchmarks.* In this subsection, we conduct a comparative analysis of the proposed approach against various approaches employed in existing literature. Tables 4.8 and 4.9 showcase the total cost and carbon emissions associated with different approaches under varied carbon pricing scenarios, respectively. We use 200 samples of historical prediction errors to construct the uncertainty set. For the oracle situation, we suppose that the load and outdoor temperature predictions are accurate and there is no uncertainty. As depicted in Tables 4.8 and 4.9, we consider the following benchmarks to handle the uncertainties inherent in TCL scheduling: scenario generation-based stochastic optimization (SG), robust

Table 4.8: Total cost of different approaches under varying carbon prices

p_{co_2} (\$/kg)	Oracle	SG	RO	CVaR-DRCC	HL-DRCC	BCD-DRCC
0.0	\$77.64	\$91.61	\$88.23	\$89.85	\$86.42	\$**85.25**
0.05	\$81.69	\$96.49	\$92.87	\$94.59	\$91.70	\$**89.96**
0.10	\$85.72	\$101.35	\$97.50	\$99.31	\$95.71	\$**94.42**

Table 4.9: Carbon emission under varying carbon prices

p_{co_2} (\$/kg)	Oracle	SG	RO	CVaR-DRCC	HL-DRCC	BCD-DRCC
0.0	82.65 kg	102.10 kg	97.45 kg	100.89 kg	98.38 kg	**96.91** kg
0.05	80.65 kg	97.17 kg	92.51 kg	94.47 kg	91.58 kg	**89.54** kg
0.10	80.54 kg	97.04 kg	92.40 kg	94.35 kg	90.68 kg	**89.37** kg

optimization (RO), and CVaR-DRCC. In the SG approach, the chance constraints are replaced with a set of sampled constraints. As all samples need to be satisfied, the SG approach can yield very conservative results. On the other hand, the inherent conservativeness of the RO approach stems from its requirement for every constraint to remain feasible across the entire range of potential values within the uncertainty set. This requirement made the RO method lack the capability to adjust the violation rate, thereby limiting the ability to optimize robustness and operational performance. Due to these major drawbacks, it can be clearly observed that both HL-DRCC and BCD-DRCC outperform SG and RO methods in almost all cases. Notably, BCD-DRCC yields the best result, with an improvement of 6.84% and 3.15% over SG and RO, respectively.

Subsequently, we evaluate the proposed methods with the CVaR approximation of DRCC. Note that for DRCC, the conservatives and robustness can be adjusted through violation rate ϵ and uncertainty set radius r. For this comparison, we first set r to $0.5\%\xi$ and ϵ to 0.1. Our observations indicate that, under the same r and ϵ, the BCD-based DRCC algorithm yields the lowest carbon emissions and overall cost. Following this, we explore varying risk levels within the CVaR-based DRCC, HL-based DRCC, and BCD-based DRCC approaches for the carbon-aware TCL scheduling

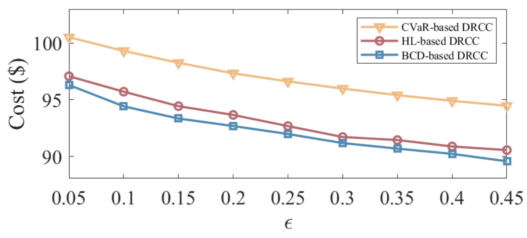


Figure 4.4: Different risk level ϵ

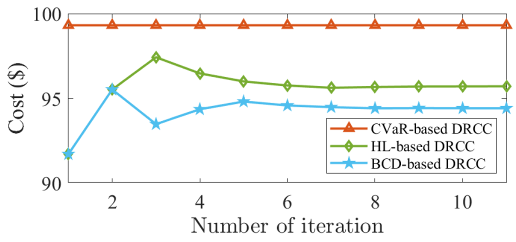


Figure 4.5: Convergence process

model. As depicted in Figure 4.4, an increase in risk level results in a decrease in cost for all three algorithms. However, this also leads to an increase in risk, indicating a trade-off between cost and risk. Moreover, at the same risk levels, the costs incurred by HL-based and BCD-based DRCC are consistently lower compared to those from the CVaR approximation-based algorithms. Specifically, at a risk level of 0.1, the HL-based and BCD-based methods respectively improved performance by 3.62% and 4.92%. The experiment results are consistent with the theoretical evidence presented in Theorem 2.

The convergence process for the scenario where the carbon price is \$0.1/kg is illustrated in Figure 4.5. The orange line represents the CVaR approximation result, which, as outlined in Algorithms 1 and 3, is used as the upper bound. A binary search is then employed to determine the optimal value. The results demonstrate that the algorithm converges within approximately 10 iterations.

2) *Impacts of carbon price and regional carbon intensity.* In this subsection, the performance of carbon-aware scheduling algorithms is assessed across different carbon pricing schemes and regional carbon intensity levels. Table 4.8 and Table 4.9 also present the relationship between total cost and carbon emissions under different carbon prices, respectively. The results reveal a clear correlation between increased carbon prices and decreased carbon emissions, indicating the effectiveness of implementing appropriate carbon pricing mechanisms in curbing environmental pollution.

Then the carbon emissions under varying regional carbon intensity levels are examined. We compare the proposed BCD-based DRCC scheduling algorithm in three regions, which are denoted as S1, S2, and S3, representing the carbon intensity in California Independent System Operator (CAISO), Northwestern Energy, and Electric Reliability Council of Texas (ERCOT), respectively. In S1, carbon intensity is comparatively low, experiencing daily fluctuations attributable to the substantial influence of renewable energy sources, like solar power, particularly in California. The period of lowest carbon intensity is observed during the afternoon. S2 is characterized by a higher carbon intensity, yet with minimal fluctuation. In S3, the carbon intensity is the highest, accompanied by significant fluctuations due to the significant share of wind/solar generation in Texas.

Table 4.10 and Table 4.11 present the total cost and carbon emissions in different regions under varying carbon prices. It can be observed that for a building located in S1, the carbon-aware scheduling algorithm can decrease carbon emissions by 7.55% when the carbon price is 0.1 $/kWh. However, in S2, the reduction is only 0.90%, due to the minimal daily fluctuation in carbon intensity. This minimal fluctuation limits the potential for reducing carbon emissions by shifting TCLs or controlling energy storage systems. In S3, carbon emissions can be reduced by 2.03%. This relatively small reduction, compared to the original daily total carbon emissions, is due to the typically high carbon intensity in S3. As a result, the effectiveness of the carbon-aware mechanism might be compromised in areas such as S2 and S3, highlighting the imperative for governments to advocate for increased adoption of renewable energy sources. This strategy not only mitigates carbon emissions itself but also motivates individuals to reduce their personal carbon footprint.

3) *Energy consumption and carbon emission of TCLs.* We set the carbon price to $0.1/kWh and the risk level to 0.1. Figures 4.6 and 4.7 depict the energy consumption of different building components and TCLs across different zones, respectively. In Figure 4.6, it is demonstrated that in a scenario sensitive to carbon emissions,

Table 4.10: Total cost under different regional carbon intensity levels

p_{co_2} (\$/kg)	S1	S2	S3
0.0	\$85.25	\$85.25	\$85.25
0.05	\$89.96	\$100.01	\$107.76
0.10	\$94.42	\$114.49	\$129.48

Table 4.11: Total emissions under different regional carbon intensity levels

p_{co_2} (\$/kg)	S1	S2	S3
0.0	96.92 kg	292.56 kg	451.63 kg
0.05	89.54 kg	290.33 kg	444.09 kg
0.10	89.37 kg	289.94 kg	442.46 kg

the ESS is programmed to initiate charging during periods of low carbon intensity, specifically between 0:00 and 2:00. Furthermore, it is designed to discharge, taking into account both the carbon intensity and the escalation in electricity prices. For instance, although the electricity prices at 14:00 and 15:00 are identical, the carbon intensity at 14:00 is 31.97% lower compared to that at 15:00. Consequently, the ESS is scheduled to discharge at 15:00, aligning with the carbon-aware objective. Figure 4.7 highlights the differential energy consumption patterns across zones, revealing a peak at 15:00, followed by a significant decrease by 16:00. This reduction is attributed to a strategic pre-cooling in Zone 1 (Z1) in anticipation of a spike in electricity prices at 16:00, aimed at cost mitigation. At this time, the reference temperature is set to 20.99°C, contrasted with a standard reference temperature of 24°C at other times. This strategy is not replicated in other zones, mainly because of a higher α value, indicating a more effective heat retention capability. Thus, zones with poor heat retention may not benefit from shifting cooling loads to reduce costs and carbon emissions.

For comparative analysis, Figures 4.8 and 4.9 illustrate the energy consumption patterns for both building components and TCLs across different zones with a carbon

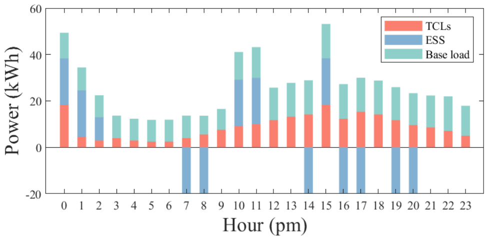


Figure 4.6: Energy consumption profiles in the TCLs under $p_{co_2} = 0.1\$/\text{kg}$

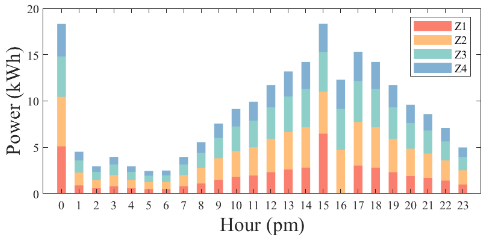


Figure 4.7: TCLs of different zones under $p_{co_2} = 0.1\$/\text{kg}$

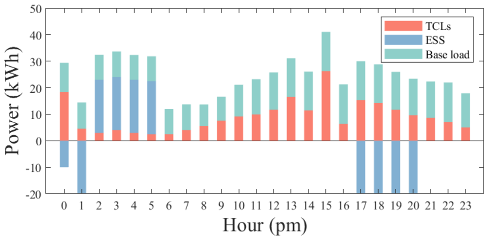


Figure 4.8: Energy consumption of different components under cost-aware scenario

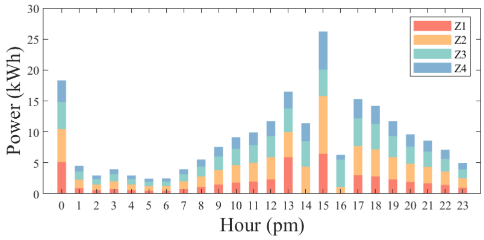


Figure 4.9: TCLs of different zones under cost-aware scenario

price of zero, which is regarded as traditional cost-aware TCL scheduling. This scenario represents a cost-aware optimization strategy that does not account for carbon emissions. Under this scenario, the ESS opts to charge during the low electricity price window of 2:00-5:00 and discharges during 17:00-20:00, disregarding carbon emissions. Conversely, with a carbon price of 0.1 \$/kWh, the ESS strategy shifts to charging between 0:00 and 2:00 and discharging at 14:00, coinciding with the increasing carbon emission and price. Zone 1 implements pre-cooling at 13:00 since the electricity price rises at 14:00. However, this pre-cooling strategy is abandoned when considering the carbon price of 0.1 \$/kWh due to the rise of carbon intensity at 13:00. This observation underscores the impact of hourly varying carbon intensity on modifying energy consumption patterns of a building to minimize costs as well as carbon emissions.

4) *Impacts of comfortable building indoor temperature range.* We then choose five comfortable ranges within the building to evaluate their impact on carbon emissions and total costs. The data presented in Table 4.12 and Table 4.13 illustrate the correlation between an expanded comfort range and a reduction in energy consumption, which subsequently leads to a substantial decrease in carbon emissions. It can be clearly observed that compared with performance benchmarks, the proposed HL-DRCC and BCD-DRCC methods can lead to up to a 5% cost and emission reductions, respectively, under any given comfortable range, for building residents.

Table 4.12: Emission under different comfortable range

Temp Range (°C)	SG	RO	CVaR-DRCC	HL-DRCC	BCD-DRCC
$[20, 24]$	97.04 kg	92.40 kg	94.35 kg	90.68 kg	89.37 kg
$[21, 25]$	93.01 kg	88.36 kg	90.31 kg	86.64 kg	85.34 kg
$[22, 26]$	88.98 kg	84.32 kg	86.28 kg	82.61 kg	81.30 kg
$[23, 27]$	84.99 kg	80.33 kg	82.29 kg	78.61 kg	77.31 kg
$[24, 28]$	81.31 kg	76.66 kg	78.61 kg	74.94 kg	73.63 kg

Table 4.13: Total cost under different comfortable range

Temp Range (°C)	SG	RO	CVaR-DRCC	HL-DRCC	BCD-DRCC
$[20, 24]$	\$101.35	\$97.50	\$99.31	\$95.71	\$94.42
$[21, 25]$	\$96.92	\$93.08	\$94.89	\$91.29	\$90.0
$[22, 26]$	\$92.51	\$88.66	\$90.47	\$86.87	\$85.58
$[23, 27]$	\$88.12	\$84.27	\$86.08	\$82.48	\$81.19
$[24, 28]$	\$83.98	\$80.12	\$81.94	\$78.34	\$77.04

4.7 Conclusion

Decarbonization of the building sector plays a pivotal role in the pursuit of climate neutrality. This paper presents a carbon-aware optimization strategy designed to enhance the efficient control of building TCLs. To mitigate uncertainties arising from outdoor temperature and base load forecast errors, we first formulate the TCL scheduling problem as a data-driven DRCC model under Wasserstein ambiguity sets, and then develop a novel reformulation strategy to recast the DRCC into a bilevel optimization structure. The proposed approach has been validated through simulation studies employing real-world data. Results have shown that the proposed approach yields a substantial reduction in regret percentage, surpassing 4.92% compared to the traditional approximation methods prevalent in existing literature for TCL control. Simulation results also demonstrate that the implementation of the carbon-aware mechanism can result in a noteworthy reduction in carbon emissions, 7.55% for our application. Furthermore, due to the alleviated conservativeness of the proposed approach, we show that significant cost and emission reductions can be achieved under any given comfortable range.

Chapter 5

24/7 Carbon-free Data Centers Operation Scheduling

5.1 Introduction

The surge in AI development and use since the release of ChatGPT in 2022, with computational power doubling approximately every 100 days and AI tasks consuming significantly more energy than traditional methods [81], is dramatically increasing electricity demand. Global data center electricity demand is projected to rise by 50% by 2027 and potentially 165% by the decade's end [79]. Consequently, carbon emissions from tech giants like Microsoft, Meta, and Google have substantially increased, largely due to AI investments [81].

To mitigate the environmental impacts of AI-driven energy demand, increasing attention is being directed toward achieving 24/7 Carbon-Free Energy (CFE). Unlike traditional renewable energy procurement, where companies purchase unbundled energy attribute certificates (EACs) to match annual consumption volumes, 24/7 CFE requires matching electricity demand with carbon-free energy generation *every hour*, in every region where operations occur. This model ensures that clean energy supply aligns with real-time consumption, reducing reliance on carbon-intensive grid electricity during periods when renewables are unavailable. Major technology companies have embraced this shift: Google has committed to operating entirely on 24/7 carbon-free energy globally by 2030, while Microsoft aims to become carbon negative by 2030 through a combination of emissions reductions and carbon removal. Meanwhile, Amazon, Meta, and Google have collectively invested in 22 GW of renewable energy capacity to support their Net Zero goals [33] by relying on locally generated renewable energy in complemantary to the electricity provided by the local grid, which in many cases is not 100% renewable powered at all hours.

However, achieving true 24/7 CFE introduces two unique challenges. First, the variability and intermittency of renewable energy sources, such as solar and wind, make it difficult to guarantee continuous clean power availability. This variability cannot be addressed solely by scaling up renewable infrastructure; instead, it necessitates a shift in how data centers manage their operations. While high-priority, user-facing services require real-time responses, a significant fraction of data center workloads, such as batch processing and AI training tasks, are latency-tolerant and governed by service level objectives (SLOs) that allow flexible execution timelines. For instance, Google reports that around 40% of tasks scheduled via its Borg system have 24-hour SLOs [165]. This latent flexibility offers a critical opportunity: by implementing emission-aware scheduling, data centers can dynamically align their power consumption with periods of renewable energy abundance, with the goal of closing the gap between supply and demand and advancing toward 24/7 CFE [33].

The second major challenge in realizing 24/7 CFE lies in managing the uncertainty inherent in forecasting future energy availability. Renewable generation is highly sensitive to weather patterns and time-of-day fluctuations, making accurate prediction difficult. Although forecasting models—typically based on historical data and machine learning techniques—can provide estimates [143], they inevitably introduce errors, especially under rapidly changing environmental conditions [191]. These prediction errors pose a significant threat to the reliability of emission-aware scheduling strategies, as misalignment between expected and actual clean energy availability can result in missed decarbonization opportunities or even system instability.

In the face of these two challenges, this paper focuses on addressing two critical research questions: 1) *How to model the energy consumption of a data center so we can match workload with low-carbon periods to achieve 24/7 CEF?* 2) *How to make robust operation and further guarantee the 24/7 renewable energy utilization utilizing the covariates information?*

In this study, we extend the application of conformal prediction by integrating it with chance-constrained programming. Specifically, quantile regression is employed to calibrate the uncertainty of the prediction model, generating finite-sample valid and conservative uncertainty sets with statistical feasibility guarantees. To further reduce the conservatism of classical pointwise conformal prediction, we propose a multi-variable conformal quantile prediction method, which improves the efficiency and practicality of uncertainty modeling. The numerical results demonstrate that the proposed methods achieve a cost reduction of up to 6.65% and a reduction of 6.96% carbon-based energy consumption compared to covariate-independent approaches, while maintaining statistical feasibility guarantees. The contributions of this study are summarized as follows:

1. We formulate a 24/7 CFE model designed specifically for optimizing data center operation scheduling, considering both flexible and inflexible load models while accounting for the dynamic carbon-based energy proportion of the power grid.

2. We propose a novel framework that integrates conformal prediction with chance-constrained programming to robustly optimize this model under renewable energy prediction uncertainty. This framework utilizes covariate information to enhance uncertainty modeling accuracy, thereby reducing operational cost and carbon-based energy of the data centers.

3. We establish theoretical guarantees that link the uncertainty sets generated by the upstream conformal prediction directly to the statistical feasibility of the downstream chance-constrained optimization problem, providing a theoretical connection between these two technologies.

The remainder of this paper is organized as follows: Section 5.3 introduces the data center power consumption model and workload scheduling model. Section 5.4 presents the chance-constrained problem, its reformulation, and the connection to conformal prediction. In Section 5.5, the conformal prediction approach for data center workload scheduling is detailed within the reformulated framework. Section 5.6 provides the numerical studies, and Section 5.7 concludes the paper.

5.2 Related Work

5.2.1 24/7 Renewable Energy Operation of Data Centers

Several approaches have been proposed to enhance the environmental and economic performance of green data centers. Ghamkhari *et al.* focuses on maximizing the profit of green data centers by incorporating and managing onsite renewable energy generators in [77]. Chen *et al.* directly addresses carbon emissions by optimizing the energy management strategies for distributed data centers operating under demand response (DR) programs in [64]. Kwon *et al.* investigate the co-optimization of server operations and power procurement in [100]. Their approach utilizes renewable energy sources and energy storage systems to facilitate power procurement and enable demand response capabilities. Kiani *et al.* introduce a model that differentiates workloads into 'green' and 'brown' categories in [98]. This decomposition allows for considering the varying costs and environmental impacts associated with green and brown energy, enabling the strategic allocation of green workloads to data centers based on the localized availability and cost of green energy, with the overar-

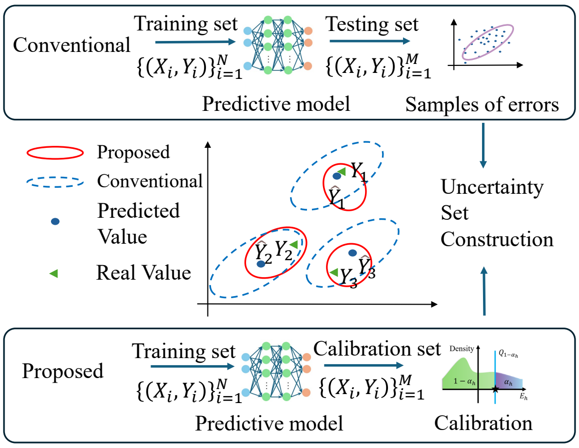


Figure 5.1: Comparison between conventional and proposed uncertainty set construction

ching goal of profit maximization. The solution for achieving 24/7 carbon-free data centers operations is first introduced in [33]. It addresses investment strategies for 24/7 carbon-free operations using energy storage, renewable energy, and workload shifting, but focuses on a long-term horizon. *Research Gap:* While existing studies provide valuable insights into cost optimization, renewable integration, and investment strategies for low-carbon data centers, there is a lack of research addressing the operational-level implementation and modulation of 24/7 carbon-free energy solutions, which is critical for achieving ambitious future data centers sustainability goals.

5.2.2 Data-Driven Approaches for Constructing Uncertainty Sets

Robust optimization (RO) is a well-established paradigm widely utilized to address inherent uncertainties, particularly those arising from intermittent renewable energy generation [182]. While commonly employed, traditional static uncertainty sets (e.g., box, budget, ellipsoidal) often yield overly conservative solutions by overemphasizing low-probability extreme scenarios. Furthermore, approaches that attempt to heuristically control uncertainty set size typically lack rigorous statistical guarantees. To mitigate this conservatism and introduce statistical rigor, research has explored integrating RO with chance-constrained programming (CCP). Specifically, data-driven CCP methods aim to construct uncertainty sets that ensure constraint satisfaction with a pre-defined probability under the worst-case distribution [67, 189], thus offering a balance between robustness and optimality.

Despite these advancements, a significant limitation of existing data-driven robust optimization approaches is that they are covariate-independent [49]. These methods typically use input features (i.e., covariates) like wind speed, irradiance, temperature, or historical generation data to train a predictive model. However, the subsequent construction of the uncertainty set (e.g., intervals for RO or error distributions for CCP) is derived solely from the prediction errors observed in a test set, without dynamically leveraging the real-time covariate information at the prediction moment. Therefore, as illustrated in Fig. 5.1, the size and shape of the generated uncertainty set often remain uniform across different new data points, despite varying underlying input conditions. This contrasts with an intuitive understanding that the level of uncertainty might change based on specific environmental or operational covariates. For the downstream optimization, the conservative uncertainty set will increase total operational costs for data centers.

Some recent efforts have sought to incorporate covariate information into the dynamic sizing of uncertainty sets [49]. However, these methods are often restricted to specific prediction models (e.g., ensemble learning), involve computationally intensive dual theory or sample-based techniques, or are limited to addressing uncertainty solely in the objective function without providing feasibility guarantees for constraints [155]. *Research Gap*: Traditional robust and data-driven stochastic optimization methods overlook the direct relationship between exogenous input features (covariates) and the structure of prediction uncertainty. This oversight results in overly conservative decisions, as they fail to dynamically adapt the uncertainty set's characteristics based on real-time contextual information. Addressing this gap requires methods that explicitly integrate covariate information into the construction of statistically rigorous and adaptive uncertainty sets for optimization.

5.3 Model Formulation

This section presents the modeling of a datacenter microgrid, followed by a description of the power consumption model of the data center, the carbon-based energy accounting model, the objective function, the energy storage system (ESS) model, and the power balance. The energy procurement decision from the utility grid and the scheduling decision for onsite generation resources are co-optimized with the workload allocation decision.

5.3.1 Data Centers Power Consumption Model

A data center typically comprises three primary components: IT equipment, refrigeration equipment, and the power distribution system [187]. Power Usage Effectiveness (PUE) is a widely adopted metric for assessing data center energy consumption. PUE is defined as the ratio of the total power consumption of the data center to the power consumed solely by the IT equipment. Following the model in [187] and [197], the total active power load of the data center can be determined as follows:

$$\begin{aligned} Q_t^{\mathrm{DC}} = &\left[P^{\mathrm{idle}} + (PUE - 1)\,P^{\mathrm{peak}}\right] A_t^{\mathrm{DC}} \\ &+ W_t^{\mathrm{DC}}\left(P^{\mathrm{peak}} - P^{\mathrm{idle}}\right)/L^{\mathrm{rate}}, \forall t \in T \end{aligned} \tag{5.1}$$

$$W_t^{DC} = \sum\nolimits_c^{\mathcal{C}} L_{c,t}^{FL} + L_t^{IFL}, \forall t \in T \tag{5.2}$$

$$A_t^{\mathrm{DC}} = A_t^{\mathrm{IFL}} + A_t^{\mathrm{FL}}, \forall t \in T \tag{5.3}$$

where P^{idle} and P^{peak} denote the power consumption of a server when it is not processing tasks and when it is actively processing tasks, respectively. W_t^{DC} denotes the total arriving workloads in the data center, $L_{j,t}^{\mathrm{FL}}$ denotes the arriving class c workloads at time t, and L_t^{IFL} represents the arriving inflexible workloads at time t. A_t^{IFL} and A_t^{FL} are the number of active servers for inflexible workloads and flexible workloads, respectively. L^{rate} is the service rate of a server. Eq. (5.2) calculates the total workload, including both flexible and inflexible workloads. Eq. (6.11) indicates the number of active servers in the data center.

5.3.2 Data Centers workload Scheduling Model

Flexible workloads model. Flexible workloads, such as big data analysis, image processing, and deep learning model training, can be deferred within a reasonable time. Following the model in [83], it is assumed that each class of workloads $c \in \mathcal{C}$ is associated with a temporal flexibility parameter $h_c \in \mathbb{Z}_{\geq 0}$, which represents the maximum allowable delay for workloads of that class. Consequently, the flexible workloads scheduling model is as follows:

$$\sum\nolimits_{\tau=1}^{t} L_{C,\tau}^{\mathrm{FL}} \leq \sum\nolimits_{\tau=1}^{t} \bar{L}_{C,\tau}^{\mathrm{FL}}, \forall t \in \mathcal{T}, \forall c \in \mathcal{C}, \tag{5.4}$$

$$\sum\nolimits_{\tau=1}^{t+h_c} L_{c,\tau}^{\mathrm{FL}} \geq \sum\nolimits_{\tau=1}^{t} \bar{L}_{c,\tau}^{\mathrm{FL}}, \forall t \in [1, |T| - h_c], \forall c \in \mathcal{C}, \tag{5.5}$$

$$\sum\nolimits_{\tau=1}^{T} L_{c,\tau}^{\mathrm{FL}} \geq \sum\nolimits_{\tau=1}^{t} \bar{L}_{c,\tau}^{\mathrm{FL}}, \forall t \in [|T| - h_c + 1, |T|], \forall c \in \mathcal{C}, \tag{5.6}$$

$$L^{\mathrm{rate}} A_t^{\mathrm{FL}} \geq \sum\nolimits_{c \in \mathcal{C}} L_{c,t}^{\mathrm{FL}}, \forall t \in \mathcal{T} \tag{5.7}$$

$$L_t^{\mathrm{FL}}, A_t^{\mathrm{FL}} \geq 0, \forall t \in \mathcal{T}. \tag{5.8}$$

Eq. (5.4) indicates that the processed flexible workload does not exceed the accumulated flexible workload up to time t. Eq. (5.5) specifies that the accumulated flexible workload must be processed within the maximum allowable delay time h_c. Eq. (5.6) suggests that all flexible workloads should be processed within a single day. It is assumed that all the flexible workloads processed in the homogeneous active servers must satisfy the total number of inflexible workloads as specified in Eq. (5.7). Eq. (5.8) ensures that the number of inflexible workloads and the corresponding active servers are positive.

Inflexible workloads model. Within data centers, there exist user-facing services, such as inference, which mandate real-time responses and require immediate processing. These are categorized as inflexible workloads, and their operation model is expressed as follows:

$$L_t^{\mathrm{IFL}} \geq \bar{L}_t^{\mathrm{IFL}}, \forall t \in \mathcal{T}, \tag{5.9}$$

$$\frac{1}{L_j^{\mathrm{rate}} - L_{j,t}^{\mathrm{IFL}} / A_{j,t}^{\mathrm{IFL}}} \leq \mathrm{C}^{\mathrm{DT}}, \forall j \in D, \forall t \in \mathcal{T}, \tag{5.10}$$

$$L_{j,t}^{\mathrm{IFL}}, A_{j,t}^{\mathrm{IFL}} \geq 0, \forall j \in D, \forall t \in \mathcal{T}, \tag{5.11}$$

$$A_{j,t}^{\mathrm{DC}} \leq A_{j,\max}^{\mathrm{DC}}, \forall j \in D, \forall t \in \mathcal{T}, \tag{5.12}$$

Eq. (5.9) denotes that the dispatched inflexible workloads at the data center must meet or exceed the arriving inflexible workloads at the front-end server. Eq. (5.10) describes that the active servers responsible for processing inflexible workloads must adhere to users' time delay requirements, as defined by the quality of service (QoS). Eq. (5.11) ensures that both the number of inflexible workloads and the corresponding active servers are positive. Eq. (5.12) restricts that the number of active servers in the first stage should not exceed the total number of servers.

2) *Grid Carbon-based Energy Accounting*

This study assumes that the data center is connected to the main power grid. The grid's power supply is derived from a mix of energy sources, including carbon-free energy such as wind, solar, and nuclear, as well as carbon-based energy such as oil, natural gas, and coal-fired power. To enable a more accurate assessment, a carbon-based energy accounting model is employed. This model quantifies the proportion of carbon-based energy per unit of electricity generated, as defined by the following equation:

$$\sum\nolimits_{t=1}^{T} CB_t \cdot P_{buy,t}, \tag{5.13}$$

where $P_{buy,t}$ is the power purchased by the data center at time t, and CB_t denotes the proportion of carbon-based energy in the grid at time t. It is further expressed as:

$$CB_t = \frac{\sum_{k\in\mathcal{E}_{CB}} M_t^k}{\sum_{k\in\mathcal{E}} M_t^k}, \tag{5.14}$$

where $\mathcal{E}$ represents the set of all energy sources (e.g., wind, solar, coal), $\mathcal{E}_{CB}$ denotes the subset of carbon-based energy sources (e.g., coal, oil, natural gas), and M_t^k is the electricity output from energy source k at time t.

CB_t reflects the average proportion of carbon-based energy in the grid at a given time and varies hourly based on the real-time contribution of each energy source. It provides a more accurate measurement of carbon emissions associated with electricity consumption. And the real-time data on the composition of energy sources is usually provided by Independent System Operators (ISOs) [6]. To achieve 24/7 CFE, data centers are required to procure Time-Based Energy Attribute Certificates (T-EACs). The total T-EACs purchased by the data center must correspond to the carbon-based energy consumed by the data center, as defined in Eq. (5.13).

Traditional Energy Attribute Certificates (EACs) match renewable generation with consumption annually, allowing claims based on total certificates purchased, irrespective of generation timing. In contrast, Time-based EACs (T-EACs) enforce stricter hourly or real-time matching, ensuring renewable energy is claimed only when generated concurrently with consumption. This granular T-EACs approach enhances the accuracy of renewable usage claims and incentivizes investments in energy storage and grid management to better synchronize supply and demand, supporting 24/7 carbon-free energy systems.

5.3.3 ESS Model

As energy storage technology has advanced, energy storage systems (ESS) have become a widely adopted and cost-effective solution for storing renewable energy [33]. Data centers usually incorporate ESS to enhance system resilience and mitigate power peaks. Moreover, ESS can support 24/7 CFE operations by storing surplus power when renewable energy generation exceeds demand and discharging stored energy during periods of insufficient renewable supply. The ESS is modeled as:

$$Q_{ess,t} = Q_{ess,t-1} + P_{ess,t}, \tag{5.15}$$

$$Q_{ess,min} \leq Q_{ess,t} \leq Q_{ess,max}, \tag{5.16}$$

$$P_{ess,min} \leq P_{ess,t} \leq P_{ess,max}, \tag{5.17}$$

where $P_{ess,min}$ and $P_{ess,max}$ are the minimum and maximum charge and discharge power of the ESS. $Q_{ess,t}$ denotes the electricity storage level of the ESS at t. $P_{ess,t}$ is charging and discharging power of ESS at time t (positive when charging and negative when discharging).

5.3.4 Chance-Constraints Problem Formulation

For the data center, the power supply must be greater than the power demand. Considering the uncertainty in renewable energy generation, the power balance for electricity is expressed as chance constraints, formulated as follows:

$$\Pr\left(P_t^{\text{Renew}} + P_t^{\text{e,grid}} \geq Q_t^{DC} + P_t^{ESS}\right) \geq 1 - \epsilon, \forall t \in \mathcal{T}, \tag{5.18}$$

where $P_t^{\text{e,grid}}$ is the purchased electricity at time t; P_t^{renew} is the renewable energy generation at time t. ϵ is the violation rate preset by the data center.

Therefore, the overall 24/7 CFE data centers operation scheduling problem is formulated as follows:

$$(\text{CCP}) \min \sum_{t=1}^{T} \lambda_t^e P_t^{\text{grid}} + \lambda^c CE, \tag{5.19}$$

$$s.t.(6.10) - (5.18) \tag{5.20}$$

where λ_t^{e} and λ^{c} are the electricity price and T-EACs price, respectively. The objective of the data center microgrid is to minimize the total operational cost, which comprises the costs of energy procurement and the costs associated with purchasing T-EACs.

5.4 Proposed Framework of Conformal Prediction-based Chance Constrained Programming

In this section, we begin by revisiting the formulation of CCP and present the fundamental procedural framework of our algorithm. Subsequently, we present the methodologies that have been developed in our research.

5.4.1 Contextual Chance-Constraints Programming

The common form of a traditional CCP typically takes the following structure:

$$(\text{T-CCP}) \text{ min } f(x); \text{ s.t. } P(g(x;\xi) \in \mathcal{A}) \geq 1-\epsilon \tag{5.21}$$

where x is the decision variable, ξ is a random variable that can be observed via a finite amount of data, $\mathcal{A}$ is the feasible region, and ϵ is the tolerance level.

In the traditional CCP, the random variable ξ is estimated using a finite set of historical data. However, this estimation process is independent of any additional features or information, which limits its modeling capability when dealing with complex distributions in practical scenarios. With the increasing availability of data, there has been a growing interest in formulating optimization under uncertainty as contextual optimization problems. These problems aim to leverage rich feature observations to improve decision-making [46]. In this context, we employ a contextual C-CCP approach [137]:

$$\text{(C-CCP)} \min f(x);\ \text{s.t.}\ P_{\xi|\psi}(g(x;\xi) \in \mathcal{A}) \geq 1-\epsilon \tag{5.22}$$

The probability is considered with respect to the conditional distribution $\xi \mid \psi$. The "contextual" decision maker has access to a vector of covariates z, which is assumed to be correlated with ξ. Consequently, this decision maker seeks to identify an optimal policy, defined as a functional $\boldsymbol{x} : \mathbb{R}^m \to \mathcal{X}$ that gets an action in $\mathcal{X}$ based on the observed realization of ψ [65].

5.4.2 The RO Reformulation of the C-CCP

The original chance constraints are intractable due to their non-differentiability and non-convexity, necessitating the use of an approximation. To address this problem, the model described in (5.21) is transformed into a robust optimization (RO) problem. This transformation also enables the direct utilization of data [48]. Robust optimization addresses uncertainty by representing it through a deterministic set, known as an uncertainty set. The RO formulation ensures that the safety condition is satisfied for any value of ξ within the uncertainty set, thereby simplifying the problem and enhancing its tractability. The RO formulation of (5.21) is as follows:

$$\text{(RO)} \min f(x)\ \text{s.t.}\ g(x;\xi) \in \mathcal{A}, \forall \xi \in \mathcal{U} \tag{5.23}$$

where $\mathcal{U} \in \Omega$ is an uncertainty set.

Theorem 5.4.1. *Any feasible solution to problem RO (5.25) is also feasible to problem T-CCP (5.21) if the following condition holds for $\mathcal{U}$:*

$$\mathbb{P}_{\xi}(\boldsymbol{\xi} \in \mathcal{U}) \geq \rho \tag{5.24}$$

Proof. Obviously, for any x feasible for (5.25), $\xi \in \mathcal{U}$ implies $g(x;\xi) \in \mathcal{A}$. Therefore, by choosing $\mathcal{U}$ that covers a $1-\epsilon$ content of ξ (i.e., $\mathcal{U}$ satisfies $P(\xi \in \mathcal{U}) \geq 1-\epsilon$), any x feasible for (5.25) must satisfy $P(g(x;\xi) \in \mathcal{A}) \geq P(\xi \in \mathcal{U}) \geq 1-\epsilon$, implying that x is also feasible for (5.21). □

In the meantime, we can address the contextual chance-constrained problem from the perspective of robust optimization. Specifically, we consider a contextual decision maker who aims to utilize side information in the design and solution of a robust optimization problem. This approach naturally leads to the following equivalent formulation of C-CCP:

$$\min\ f(x) \text{ s.t. } g(x;\xi) \in \mathcal{A}, \forall \xi \in \mathcal{U}(\psi) \tag{5.25}$$

where $\mathcal{U}(\psi)$ represents an uncertainty set that is constructed to encompass, with high probability, the realization of ξ given the observation of ψ. The proposed approach is data-driven, as it utilizes historical observations of joint realizations of ψ and ξ in the design of the contextual robust optimization problem.

Following the schedule outlined in Theorem 1, we can conclude the following:

Theorem 5.4.2. *Any feasible solution to the robust optimization problem RO (5.25) is also feasible for the contextual chance-constrained problem C-CCP (5.21), if the following condition holds for $\mathcal{U}$:*

$$\mathbb{P}_{\xi|\psi}(\boldsymbol{\xi} \in \mathcal{U}(\psi)) \geq \rho \tag{5.26}$$

This leads to the challenge of determining the uncertainty set when side information ψ is available. To address this issue, we introduce the idea of uncertainty quantification and conformal prediction.

5.4.3 Uncertainty Quantification and Conformal Prediction

Conformal prediction is a statistical and distribution-free technique that quantifies uncertainty for any black-box predictive model. It was first introduced by Vork [168] as a postprocessing method and can be applicable for both classification and regression tasks. Consider a dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$, where each data point (X_i, Y_i) comprises a covariate vector X_i and a corresponding response Y_i. Conformal prediction aims to find a prediction band $C(X_{n+1})$ for a new observation X_{n+1}. This prediction band is designed to contain the true, but unknown response Y_{n+1} with a user-specified probability under the assumption of exchangeability ($\mathcal{D}$ and (X_{n+1}, Y_{n+1}) are exchangeable, or more weak i.i.d). Specifically, for a significance level α, we aim to construct $C(X_{n+1})$ such that:

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha, \tag{5.27}$$

Building on the foundation of conformal prediction, we now focus on the practical applications in real-world scenarios. It's widely acknowledged that the full conformal prediction is computationally expensive as the construction of the prediction set needs to invert the score function to identify possible values $y \in \mathcal{Y}$ for the response Y_{n+1} that agree (or conform) with the trends observed in the available data [41].

To address this, split conformal prediction was developed to offer a computationally efficient alternative by dividing the data to avoid the need of retraining the conformal score function. While our proposed approach can apply to both versions of conformal prediction, in the interest of space we will restrict our attention to split conformal prediction and refer the reader to [41] for a more detailed comparison between these two methods.

Split conformal prediction begins by partitioning the training data $\mathcal{D}$ into two disjoint sets: a proper training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$. First, we train a regression model $\hat{\mu}(x)$ on the proper training set using any regression algorithm $\mathcal{A}$:

$$\hat{\mu}(x) \leftarrow \mathcal{A}\left(\{(X_i, Y_i) : i \in \mathcal{I}_1\}\right). \tag{5.28}$$

Next, we calculate the calibration residuals R_i on the calibration set, which measure the prediction error:

$$R_i = \left|Y_i - \hat{\mu}\left(X_i\right)\right|, \quad i \in \mathcal{I}_2 \tag{5.29}$$

To determine the width of our prediction band, we compute the $(1-\alpha)$-quantile of these residuals. Specifically, we use a slightly conservative quantile $Q_{1-\alpha}\left(R, \mathcal{I}_2\right)$ to ensure coverage:

$$Q_{1-\alpha} := quantile(\{R_i : i \in \mathcal{I}_2\}, (1-\alpha)\left(1 + 1/\left|\mathcal{I}_2\right|\right)). \tag{5.30}$$

Finally, for a new data point X_{n+1}, the split conformal prediction interval is constructed as:

$$C\left(X_{n+1}\right) = \left[\hat{\mu}\left(X_{n+1}\right) - Q_{1-\alpha},\ \hat{\mu}\left(X_{n+1}\right) + Q_{1-\alpha}\right]. \tag{5.31}$$

This interval is guaranteed to satisfy the coverage property in equation (5.27). In summary, we have the following Theorem:

Theorem 5.4.3. *If dataset $\mathcal{D} = (X_i, Y_i)$ for any $i \in \{1, \ldots, n\}$ are exchangeable. For a new exchangeable pair (X_{n+1}, Y_{n+1}), conformal prediction construct a prediction set $C(X_{n+1})$ with the following marginal coverage guarantee:*

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha, \tag{5.32}$$

where α is predefined confidence level. Note that if R_i has a continuous joint distribution, then equation (5.32) has a tigher bound:

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1}, \tag{5.33}$$

Proof for lower bound. Obvious, the lower bound $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) = P(R_{n+1} \leq Q_{1-\alpha}) \geq 1-\alpha$ holds because $Q_{1-\alpha}$ is defined as the $(1-\alpha)$ quantile of calibration residuals. By definition, this quantile ensures that at least a $(1-\alpha)$ proportion of calibration residuals (and thus, in expectation, future residuals) are smaller, guaranteeing the desired coverage. □

Proof for upper bound. If R_i have a continuous joint distribution, then with probability 1, all scores are distinct. In this case, assume $k = \lceil (1-\alpha)(n+1) \rceil$, then $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) = P(R_{n+1} \leq Q_{1-\alpha}) = P(R_{n+1} \leq S_k) = \frac{k}{n+1} = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \leq 1-\alpha+\frac{1}{n+1}$. □

Thus, for new data sample X_{n+1}, the interval Y_{n+1} can be interpreted as the uncertainty set $\mathcal{U}(X_{n+1})$ described in Section 5.4.2. Consequently, the problem can be addressed using conventional robust optimization techniques [47]. Note that in this research, we consider the box-uncertainty set. It can also be extended to the elliptical-uncertainty set, we refer [180] for more details. A key limitation of this basic split conformal interval is that it applies a uniform interval $2Q_{1-\alpha}$ to all the samples without accounting for the varying difficulty of each sample. This motivates the need for more adaptive approaches, such as variable-width conformal prediction methods.

5.5 The Proposed Approach

5.5.1 Conformal Quantile Regression

Unlike the uniform width interval of standard split conformal prediction, Conformal Quantile Regression (CQR) leverages separately trained lower and upper quantile predictions to adapt the interval width based on the input X_{n+1}. This approach aims to provide more informative and efficient prediction intervals while maintains

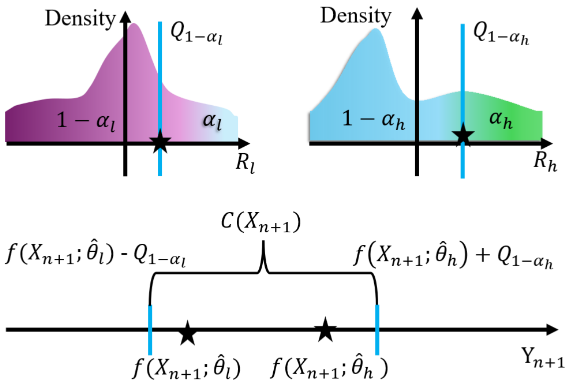


Figure 5.2: The process getting prediction interval $C(X_{n+1})$.

the same statistical coverage property. Fig. 5.2 illustrate the process of getting X_{n+1}.

Suppose $\mu(X;\hat{\theta}_l)$ and $\mu(X;\hat{\theta}_h)$ be the lower and upper quantile regression models trained on the proper training set $\mathcal{I}_1$. Compared with equation (5.29), CQR based method build the calibration residuals R_i^l and R_i^h by lower bound and upper bound seperately. Let's define the nonconformity scores on the calibration set $\mathcal{I}_2$ as:

$$R_i^l = \mu(X_i;\hat{\theta}_l) - Y_i, \quad i \in \mathcal{I}_2 \tag{5.34}$$

$$R_i^h = Y_i - \mu(X_i;\hat{\theta}_h), \quad i \in \mathcal{I}_2 \tag{5.35}$$

Let $Q_{1-\alpha_l}$ be the $(1-\alpha_l)$-th quantile of $\{R_i^l : i \in \mathcal{I}_2\}$ adjusted with the factor $(1+1/|\mathcal{I}_2|)$, and $Q_{1-\alpha_h}$ be the $(1-\alpha_h)$-th quantile of $\{R_i^h : i \in \mathcal{I}_2\}$ adjusted with the factor $(1+1/|\mathcal{I}_2|)$. Construct the CQR prediction interval for a new data point X_{n+1} as:

$$C(X_{n+1}) = \left[\mu(X_{n+1};\hat{\theta}_l) - Q_{1-\alpha_l}, \mu(X_{n+1};\hat{\theta}_h) + Q_{1-\alpha_h}\right].$$

Then, for a given significance level $\alpha = \alpha_l + \alpha_h$, the CQR prediction interval satisfies the marginal coverage guarantee:

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha. \tag{5.36}$$

Theorem 5.5.1. *If dataset $\mathcal{D} = (X_i, Y_i)$ for any $i \in \{1,\ldots,n\}$ are exchangeable. For a new exchangeable pair (X_{n+1}, Y_{n+1}), conformal quantile regression constructs a* *prediction set $C(X_{n+1})$ with the following marginal coverage guarantee:*

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha_l - \alpha_h, \tag{5.37}$$

Proof. By construction, the lower quantile residual $Q_{1-\alpha_l}$ and the upper quantile residual $Q_{1-\alpha_h}$ satisfies following properties:

$$\mathbb{P}\big(Y_{n+1} \geq \mu(X_{n+1}; \hat{\theta}_l) - Q_{1-\alpha_l}\big) \geq 1 - \alpha_l, \tag{5.38}$$

$$\mathbb{P}\big(Y_{n+1} \leq \mu(X_{n+1}; \hat{\theta}_h) + Q_{1-\alpha_h}\big) \geq 1 - \alpha_h. \tag{5.39}$$

Then prediction interval $C(X_{n+1})$ covers Y_{n+1} if both inequalities hold. Let $A = \{Y_{n+1} \geq \mu_l - Q_{1-\alpha_l}\}$ and $B = \{Y_{n+1} \leq \mu_h + Q_{1-\alpha_h}\}$. Using the union bound on their complements, we can get:

$$\begin{aligned}
\mathbb{P}(A \cap B) &= 1 - \mathbb{P}(A^c \cup B^c) && (5.40)\\
&\geq 1 - (\mathbb{P}(A^c) + \mathbb{P}(B^c)) && (5.41)\\
&\geq 1 - (\alpha_l + \alpha_h) && (5.42)
\end{aligned}$$

□

The CQR method achieves valid marginal coverage guarantees while adapting interval widths to input-dependent uncertainty through quantile regression. It leverages data-dependent adjustments to produce narrower intervals in regions of low uncertainty and wider intervals where variability is higher. This adaptability often leads to tighter, more informative prediction intervals without sacrificing coverage.

5.5.2 Multi-Variable Conformal Quantile Prediction

In many real-world applications, the response variable Y is not a scalar but a vector, i.e., $Y \in \mathbb{R}^d$. In this section, we extend conformal prediction to multi-variable outputs and demonstrate two distinct approaches to multi-variable conformal quantile prediction: one focusing on marginal coverage across all output variables and another aiming for individual coverage for each output variable.

Average Marginal Multi-Variable CQR

In this section, we consider a weaker notion of marginal coverage for multi-variable output, which we term Average Marginal Multi-Variable CQR (AMV-CQR). Here, we aim to ensure that the average coverage across all output dimensions is at least $1 - \alpha$. Formally, we want to satisfy:

$$\frac{1}{d} \sum_{j=1}^{d} \mathbb{P}(Y_{n+1}^{j} \in C^{j}(X_{n+1})) \geq 1 - \alpha. \tag{5.43}$$

where $Y_{n+1} = [Y_{n+1}^{1}, \ldots, Y_{n+1}^{d}]^{\top}$ and Y_{n+1}^{j} is the j-th output variable. $C(X_{n+1}) = C^{1}(X_{n+1}) \times \ldots \times C^{d}(X_{n+1})$ is a hyper-rectangular prediction region, with $C^{j}(X_{n+1})$ being the prediction interval for the j-th output variable.

Note that we need a multivariate output model for AMV-CQR which can generate lower and upper bounds for all response variables. For brevity, we share the lower and upper bound models for multivariate prediction and only change the heads of different outputs. The details of the process are described in Algorithm 6.

Theorem 5.5.2. *If dataset $\mathcal{D} = (X_i, Y_i)$ for any $i \in \{1, \ldots, n\}$ are exchangeable, and $Y_i \in \mathbb{R}^d$. For a new exchangeable pair (X_{n+1}, Y_{n+1}), the AMV-CQR algorithm constructs prediction intervals $C^{1}(X_{n+1}), \ldots, C^{d}(X_{n+1})$ such that the average marginal coverage guarantee is satisfied:*

$$\frac{1}{d} \sum_{j=1}^{d} \mathbb{P}(Y_{n+1}^{j} \in C^{j}(X_{n+1})) \geq 1 - \alpha.$$

Proof. The result follows by a simple extension of the univariate case. Since the calibration residuals across both indices $j \in \mathcal{J}$ are exchangeable, we can flatten them and apply the standard conformal procedure. Hence, as in Theorem 5.4.3, the lower bound and upper bound of the prediction set hold trivially. $\square$

Algorithm 6 Average Marginal Multi-Variable CQR (AMV-CQR)

Input: $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, $Y_i \in \mathbb{R}^d$, significance α, quantiles α_l, α_h, where $\alpha_l + \alpha_h = \alpha$, the index set $\mathcal{J} = \{1, 2, \ldots, d\}$.

Output: Prediction intervals $C^1(X_{n+1}), \ldots, C^d(X_{n+1})$.

Split $\mathcal{D}$ into proper training set $\mathcal{I}_1$ and calibration set $\mathcal{I}_2$.

Train lower and upper quantile models $\mu(\cdot;\theta_l), \mu(\cdot;\theta_h)$ on proper training set $\mathcal{I}_1$.

for $j \in \mathcal{J}, i \in \mathcal{I}_2$ **do**

Calculate residuals:

$$R_i^{l,j} = \mu^j(X_i;\theta_l) - Y_i^j, R_i^{h,j} = Y_i^j - \mu^j(X_i;\theta_l).$$

Compute quantiles:

$$\beta_l = (1-\alpha_l)(1 + 1/(|\mathcal{I}_2|d)), \beta_h = (1-\alpha_h)(1 + 1/(|\mathcal{I}_2|d)),$$

$$\begin{aligned} Q_{1-\alpha_l} &= quantile\left(\{R_i^{l,j} : i \in \mathcal{I}_2,\ j \in \mathcal{J}\},\ \beta_l\right), \\ Q_{1-\alpha_h} &= quantile\left(\{R_i^{h,j} : i \in \mathcal{I}_2,\ j \in \mathcal{J}\},\ \beta_h\right). \end{aligned}$$

end for

for $j \in \mathcal{J}$ **do**

Construct prediction interval X_{n+1}^j:

$$C^j(X_{n+1}) = \left[\mu^j(X_{n+1};\theta_l) - Q_{1-\alpha_l}, \mu^j(X_{n+1};\theta_h) + Q_{1-\alpha_h}\right]$$

end for

Return the prediction intervals $C^1(X_{n+1}), \ldots, C^d(X_{n+1})$.

Individual Multi-Variable CQR

In contrast to average marginal coverage, individual coverage focuses on ensuring coverage for each output variable separately. For a given significance level α, we aim to construct prediction intervals $C^j(X_{n+1})$ for each output dimension $j \in \{1, \ldots, d\}$ such that for each j:

$$\mathbb{P}(Y^j_{n+1} \in C^j(X_{n+1})) \geq 1 - \alpha. \tag{5.44}$$

Note that individual coverage is a stronger guarantee than average marginal coverage. If individual coverage holds, average marginal coverage automatically holds for the same α. In this section, we propose the Individual Multi-Variable CQR (IMV-CQR) algorithm to achieve individual coverage. The specifics of the process are outlined in Algorithm 2.

Theorem 5.5.3. *If dataset $\mathcal{D} = (X_i, Y_i)$ for any $i \in \{1, \ldots, n\}$ are exchangeable, and $Y_i \in \mathbb{R}^d$. For a new exchangeable pair (X_{n+1}, Y_{n+1}), the IMV-CQR algorithm constructs prediction intervals $C^1(X_{n+1}), \ldots, C^d(X_{n+1})$ such that for each dimension $j \in \{1, \ldots, d\}$, the individual coverage guarantee is satisfied:*

$$\mathbb{P}(Y^j_{n+1} \in C^j(X_{n+1})) \geq 1 - \alpha. \tag{5.45}$$

Proof. The result also follows by a simple extension of the univariate case. Since the calibration residuals across both indices $j \in \mathcal{J}$ are exchangeable, we apply the standard conformal procedure to each output seperately. Hence, as in Theorem 5.4.3, the lower bound and upper bound of the prediction set hold trivially. □

5.5.3 Comparing AMV-CQR versus IMV-CQR

AMV-CQR targets a collective or average marginal coverage guarantee, which is formalized as follows:

$$\frac{1}{d} \sum_{j=1}^{d} \mathbb{P}\Big(Y^j_{n+1} \in C^j(X_{n+1})\Big) \geq 1 - \alpha. \tag{5.46}$$

Algorithm 7 Individual Multi-Variable Conformal Quantile Regression (IMV-CQR)

Input: Training data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ where $Y_i \in \mathbb{R}^d$, significance α, quantiles α_l, α_h, where $\alpha_l + \alpha_h = \alpha$, the index set $\mathcal{J} = \{1, 2, \ldots, d\}$. New data point X_{n+1}.

Output: Prediction intervals $C^1(X_{n+1}), \ldots, C^d(X_{n+1})$.

1: Split $\mathcal{D}$ into proper training set $\mathcal{I}_1$ and calibration set $\mathcal{I}_2$.

2: Train lower and upper quantile models $\mu(\cdot;\theta_l), \mu(\cdot;\theta_h)$ on proper training set $\mathcal{I}_1$.

3: **for** $j \in \mathcal{J}, i \in \mathcal{I}_2$ **do**

4: Calculate residuals:

$$R_i^{l,j} = \mu^j(X_i;\theta_l) - Y_i^j, R_i^{h,j} = Y_i^j - \mu^j(X_i;\theta_l).$$

5: **end for**

6: **for** $j \in \mathcal{J}$ **do**

7: Compute quantiles:

$$\beta_l = (1-\alpha_l)(1+1/|\mathcal{I}_2|), \beta_h = (1-\alpha_h)(1+1/|\mathcal{I}_2|)$$

$$Q_{1-\alpha_l}^j = quantile\left(\{R_i^{l,j} : i \in \mathcal{I}_2\},\ \beta_l\right)$$
$$Q_{1-\alpha_h}^j = quantile\left(\{R_i^{h,j} : i \in \mathcal{I}_2\ \},\ \beta_h\right)$$

8: **end for**

9: **for** $j \in \mathcal{J}$ **do**

10: Construct prediction interval for X_{n+1}^j:

$$C^j(X_{n+1}) = \left[\mu^j(X_{n+1};\hat{\theta}_l) - Q_{1-\alpha_l}^j, \mu^j(X_{n+1};\hat{\theta}_h) + Q_{1-\alpha_h}^j\right]$$

11: **end for**

12: Return the prediction intervals $C^1(X_{n+1}), \ldots, C^d(X_{n+1})$.

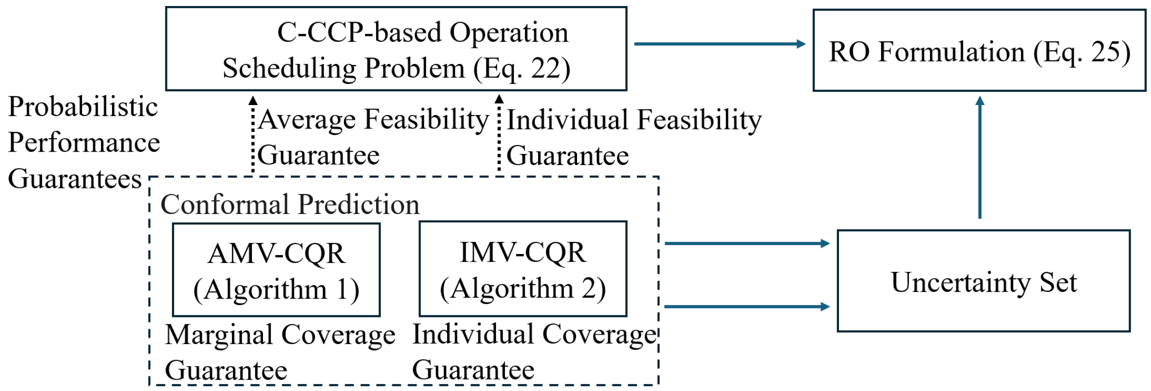


Figure 5.3: The overall process

In practice, this means that across all d output variables the mean coverage meets the desired level $1-\alpha$, even if some variables may be slightly undercovered provided that others are overcovered. This relaxed requirement may yield tighter overall prediction sets and is well suited in applications where only the aggregate reliability is of interest or when the outputs exhibit strong correlations.

In contrast, IMV-CQR enforces an individual coverage guarantee for each component $j \in \{1, \ldots, d\}$, as described by

$$\mathbb{P}\Big(Y_{n+1}^{j} \in C^{j}(X_{n+1})\Big) \geq 1-\alpha. \tag{5.47}$$

Such a guarantee is inherently stronger and more desirable when each individual prediction is required to meet a strict reliability standard. However, the additional calibration burden makes IMV-CQR more conservative in practice. Because calibration is performed for each output channel separately, the method may yield wider intervals, especially when the available calibration sample size is limited. Therefore, practitioners should balance the need for guaranteed per-coordinate coverage with the potential trade-off of less sharp intervals.

Building upon Theorem 5.4.2 and leveraging Eqs. (5.46) and (5.47), we establish that the uncertainty sets generated by our conformal prediction-based methods directly translate to statistical feasibility guarantees for the original C-CCP problem when used within the reformulated problem (5.25). This implies that using uncertainty sets from AMV-CQR yields an average statistical feasibility guarantee for C-CCP, while uncertainty sets derived from IMV-CQR ensure an individual statistical feasibility guarantee. Fig. 5.3 provides a summary of this overall methodology and the relationships between different components.

5.6 Numerical studies

5.6.1 Data Description and Experimental Settings

In this section, comprehensive numerical studies are conducted to validate the performance of the proposed conformal prediction-based chance-constraints algorithms and to analyze the sensitivity of key parameters. The parameters required for the components in the datacenter are listed in Table 1, which are set based on the research papers by [187] and [197]. All simulations were implemented using Python on a desktop computer equipped with an Intel i9 2.4GHz CPU and 32GB RAM. The optimization problem is solved using Gurobi 11.0.3.

The hourly PV generation data are generated based on historical weather data as described by [153]. The electricity prices from the utility are presented in Figure 5.4, and hourly carbon-based energy proportion (CBEP) data are obtained from Electricity Map [11]. The distribution of inflexible workloads is normalized according to the historical workload provided by [119]. Additionally, three types of flexible jobs are assumed in the experiments based on [104].

In our numerical studies, we develop distinct neural network models to predict day-ahead photovoltaic (PV) power generation. We use data from January 1, 2010, to December 31, 2019, comprising a total of 3,285 samples. The dataset is evenly split into two segments: 50% for training and 50% for testing. Specifically, we utilize two

Table 5.1: Relevant parameters for datacenter

Parameters	Values	Parameters	Values
P^{idle} (kW)	0.1	P^{peak} (kW)	0.2
PUE	1.4	A_{max}	8000
$P^{\text{ess}}_{\text{max}}$ (kW)	80	$Q_{ess,max}$ (kWh)	500
L^{rate}	3	TD (h)	[2,5,7]
C^{DT}(s)	0.5	P^{grid}_{max}	1000kW

Table 5.2: The hyper-parameters for quantile prediction

Hyper-parameters	Values	Hyper-Parameters	Values
Optimizer	AdamW	Active function	ReLU
No. of epochs	300	Hidden layers	[256,256]
No. of layers	3	Learning rate	1e-3
Quantile-low	0.1	Quantile-high	0.9

widely adopted models: the multi-layer perceptron (MLP) and the one-dimensional convolutional neural network (1D-CNN). These models take as input the PV generation data from the previous 24 hours, the predicted day-ahead surface irradiance data (weighted by land area), and temporal features such as the day of the year. Using this input, the models forecast PV generation for the following day. Additional model parameters are provided in Table 5.2.

5.6.2 Competing Methods

The methods proposed in Sections III and IV are referred to as AMV-CQR and IMV-CQR. To evaluate the performance of these approaches, they are compared against other commonly used methods for decision-making under uncertainty:

1) *Data-drivn Chance-Constrained Optimization (CCO)*: This method employs Conditional Value-at-Risk (CVaR) as an approximation, which is widely used in data-driven scenarios due to its convexity and computational tractability [189].

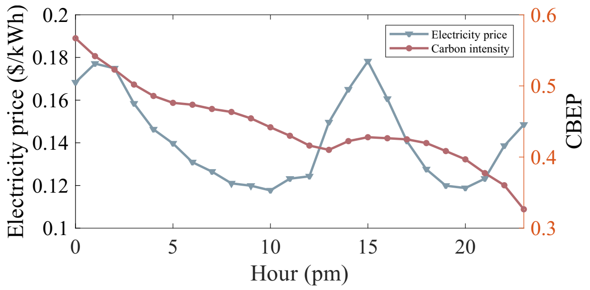


Figure 5.4: Day-ahead carbon-based energy proportion and electricity price

$1-\alpha$	RO-A	CCO	DRCCO	AMV-Point	IMV-Point	**AMV-CQR**	IMV-CQR
0.8	3449.49	3424.95	3435.72	3224.68	3288.16	**3213.32**	3276.53
0.85	3476.73	3430.78	3441.56	3252.29	3306.28	**3237.16**	3296.54
0.9	3505.23	3437.65	3448.43	3288.94	3330.01	**3272.28**	3321.80
0.95	3541.06	3447.07	3457.85	3341.19	3369.76	**3326.67**	3366.89

Table 5.3: Total cost of the datacenter using MLP prediction model with different methods and coverage levels ($)

2) *Data-Driven Distributionally Robust Chance-Constrained Optimization (DRCCO)*: This approach constructs an ambiguity set of plausible probability distributions from historical data, where the set is defined using the ∞-Wasserstein distance [94]. The objective is to identify an optimal solution under the worst-case distribution within this set.

3) *Robust Optimization with Adjusted Uncertainty Set (RO-A)*: This approach utilizes a traditional box-shaped uncertainty set, with its size adjusted to correspond to the sample coverage based on the violation rate [47].

4) *Point-Prediction-Based Conformal Prediction (AMV-Point and IMV-Point)*: This approach adheres to the proposed CP-based framework; however, in this method, the prediction and quantile regression are conducted separately. Therefore, the conformal prediction intervals for the upper and lower bounds are identical [102].

$1-\alpha$	RO-A	CCO	DRCCO	AMV-Point	IMV-Point	**AMV-CQR**	IMV-CQR
0.8	8490.36	8455.52	8481.53	7907.31	8066.81	**7881.60**	8038.85
0.85	8556.48	8470.09	8496.09	7974.70	8111.44	**7939.69**	8088.43
0.9	8625.42	8487.76	8513.76	8064.33	8169.72	**8025.41**	8150.57
0.95	8711.90	8511.50	8537.51	8192.42	8268.60	**8158.45**	8261.78

Table 5.4: Carbon-based energy consumption using MLP prediction model with different methods and coverage levels (kWh)

5.6.3 Experiments

1) *Comparing with performance benchmarks.* In this subsection, we perform a comparative analysis of the proposed approach against several methods commonly used in the existing literature. Tables 5.3 and 5.4 present the total cost and carbon-based energy consumption, respectively, for different approaches across various coverage levels. The conformal prediction model utilized in this analysis is the MLP Model. The results demonstrate that the AMV-CQR and IMV-CQR methods achieve substantial cost reductions relative to alternative approaches, including RO, data-driven CCO, and DRCCO, with maximum reductions of 6.65%, 4.81%, and 5.11%, respectively, when $1-\alpha = 0.9$. Comparable improvements are observed in terms of carbon-based energy consumption, with reductions of 6.96%, 5.45%, and 5.74%, respectively. The conservativeness exhibited by conventional methods can be attributed to their reliance on restrictive assumptions regarding the structure of the uncertainty set, as well as their inability to integrate covariate information (e.g., features of the prediction model such as weather conditions, weekday, and month). These limitations underscore the advantages of the proposed approaches, which effectively leverage contextual information to enhance predictive performance and decision-making efficiency.

The results also indicate that the CQR-based methods outperform the point-based conformal prediction methods AMV-Point and IMV-Point. This can be attributed to the fact that, in point-based conformal prediction, prediction and quantile regression are conducted separately, resulting in identical conformal prediction intervals for the

$1-\alpha$	RO-A	CC	DRCCO	AMV-Point	IMV-Point	**AMV-CQR**	IMV-CQR
0.80	0.20	0.03	0.03	0.09	0.13	0.10	0.13
0.85	0.15	0.02	0.02	0.07	0.10	0.08	0.10
0.90	0.10	0.01	0.01	0.05	0.07	0.06	0.07
0.95	0.05	0.01	0.00	0.03	0.04	0.03	0.04

Table 5.5: Probability of constraint violation in the chance-constraints for MLP prediction model under different methods

upper and lower bounds.

Table 5.5 displays the violation rates for different methods.The results confirm that all methods maintain statistical feasibility across all tested confidence levels. However, the specific values highlight trade-offs in conservatism. CVaR-CC exhibits very low violation rates (e.g., 0.01 at $1-\alpha=0.9$), indicating a highly conservative approach that secures feasibility but likely compromises solution quality. DRCCO is even more conservative than CCO because it is under the worst-case situation. While RO-A meets its target violation rate precisely, its covariate-independent uncertainty set remains static and does not adapt to real-time variations in prediction uncertainty, resulting in higher costs. In contrast, the proposed AMV-CQR and IMV-CQR methods satisfy statistical feasibility guarantees by keeping violation rates below α (e.g., 0.03 and 0.04 at $1-\alpha=0.95$). By leveraging covariate information, they produce less conservative uncertainty sets, thereby reducing costs and carbon-based energy consumption while preserving feasibility.

We also evaluated different quantile regression models for conformal prediction. Tables 5.6 and 5.7 present total cost and carbon-based energy consumption, respectively, when using a 1DCNN model for prediction across various coverage levels $1-\alpha$. Comparative analysis with an MLP-based conformal prediction model revealed no significant overall performance difference; each model excelled in certain cases. This suggests that both 1DCNN and MLP sufficiently leverage covariate information to characterize prediction errors in this scenario.

$1-\alpha$	RO-A	DD-CC	DRCCO	AMV-Point	IMV-Point	**AMV-CQR**	IMV-CQR
0.8	3449.49	3424.95	3435.72	3224.68	3286.64	**3218.48**	3280.06
0.85	3476.73	3430.78	3441.56	3253.22	3305.96	**3241.74**	3299.66
0.9	3505.23	3437.65	3448.43	3287.71	3330.44	**3274.64**	3327.20
0.95	3541.06	3447.07	3457.85	3339.55	3366.16	**3328.24**	3365.87

Table 5.6: Total cost of the datacenter using 1DCNN prediction model with different methods and coverage levels ($)

$1-\alpha$	RO-A	CC	DRCCO	AMV-Point	IMV-Point	**AMV-CQR**	IMV-CQR
0.8	8490.36	8455.52	8481.53	7907.87	8063.81	**7894.25**	8047.22
0.85	8556.48	8470.09	8496.09	7977.50	8111.18	**7950.95**	8095.45
0.9	8625.42	8487.76	8513.76	8061.80	8171.29	**8031.18**	8163.57
0.95	8711.90	8511.50	8537.51	8188.86	8260.56	**8162.26**	8259.00

Table 5.7: Carbon-based energy consumption for 1DCNN prediction model with different methods and coverage levels (kWh)

2) *Impacts of T-EACs price.* In this subsection, the performance of carbon-aware scheduling algorithms is assessed across different T-EACs pricing schemes. Table 5.8 and Table 5.9 also present the relationship between total cost and carbon-based energy consumption under different T-EACs prices, respectively. The results reveal a clear correlation between increased T-EACs prices and decreased carbon-based energy consumption, indicating the effectiveness of implementing appropriate T-EACs pricing mechanisms in curbing environmental pollution. It can be clearly observed that compared with performance benchmarks, the proposed AMV-CRQ and IMV-CRQ methods can lead to about 6% cost and carbon-based energy consumption, respectively, under any given T-EACs price.

3) *Energy consumption of different flexible workload*

The carbon price is set to $0.1/kWh, and the risk level is configured to 0.1. Figure 5.6 illustrates the power consumption of the workload under the carbon-aware scenario, while Fig. 5.5 presents the original workload arrival pattern. The results

λ_c	RO-A	CCO	DRCCO	AMV-Point	IMV-Point	**AMV-CQR**	IMV-CQR
0.00	2638.12	2584.21	2592.40	2477.95	2508.49	**2465.18**	2502.19
0.05	3072.36	3011.66	3021.14	2884.12	2919.93	**2869.41**	2912.68
0.10	3505.23	3437.65	3448.43	3288.94	3330.01	**3272.28**	3321.80
0.20	4365.37	4284.27	4297.65	4092.97	4144.58	**4072.41**	4134.46

Table 5.8: Total cost of the datacenter with different methods and carbon prices ($)

λ_c	RO-A	CCO	DRCCO	AMV-Point	IMV-Point	**AMV-CQR**	IMV-CQR
0.00	8687.14	8555.26	8580.63	8125.69	8231.08	**8086.77**	8211.93
0.05	8681.87	8544.22	8570.22	8120.78	8226.16	**8081.85**	8207.01
0.10	8625.42	8487.76	8513.76	8064.33	8169.72	**8025.41**	8150.57
0.20	8579.65	8446.83	8472.84	8018.61	8123.99	**7979.68**	8104.85

Table 5.9: Carbon-based energy consumption with different methods and T-EACs prices (kWh)

indicate that, under the carbon-aware scenario, the flexible load is shifted to hours 5–12 (when electricity prices are low) and hours 18–22 (when electricity prices and carbon-based energy proportion are low) for execution. For flexible workload 3, which has a longer time flexibility, the majority of processing occurs during periods of low electricity prices, high renewable generation, or low carbon intensity. In contrast, flexible workload 1, with a shorter delay tolerance of 2 hours, is observed to still require processing during periods of high carbon-based energy proportion and electricity price.

5.7 Conclusion

Decarbonizing data centers is essential for achieving climate neutrality. In this study, a carbon-conscious optimization strategy is devised to enable the efficient workload scheduling of data centers. The proposed model is designed to minimize overall energy expenses while simultaneously reducing carbon-based energy consumption to-

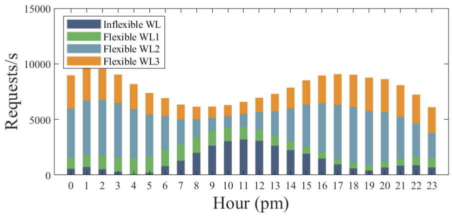


Figure 5.5: Pattern of different classes of workloads

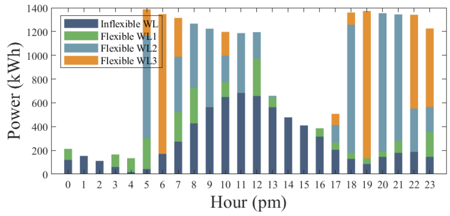


Figure 5.6: Power consumption of different kinds of workloads ($\lambda_c = \$0.1/\text{kWh}$)

ward 24/7 CFE. To address the uncertainties inherent in renewable generation predictions, a conformal prediction-based optimization framework is developed, providing statistical feasibility guarantees. The approach effectively manages uncertainties and ensures reliable and robust operation by integrating covariate information from the prediction model. The proposed framework is validated through numerical studies, demonstrating improved control over violation rates and achieving lower costs. Furthermore, the proposed approach achieves a cost reduction of over 6% in data center workload scheduling compared to traditional methods.

Chapter 6

Data Centers Workload Planning with Small Modular Reactors

6.1 Introduction

We have seen a rapid growth of AI applications in the past decade [101]. To run these applications, significant computing resources are needed [112]. GPUs have become more powerful, and increasingly energy consuming [188]. 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 [128].

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 [110, 84, 162]. 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 [86, 87]. 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 [96]; Meta issued a request for proposals in 2024 for up to 4 GW of new nuclear capacity to come online in the early 2030s [32]; and Microsoft has acquired and reopened the nuclear power plant at Three Mile Island [125].

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 [123]. In multi-energy systems, NPPs can co-generate heat and electricity for district use [140, 138], while other studies explore NPPs in energy markets for providing electricity, reserve services, and thermal products [139, 158]. 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 [151].

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 [76, 38]. 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 [93]. Moreover, excess energy generation from the LF-SMR can result in inefficient curtailment penalties, as unused nuclear energy imposes grid stability and economic challenges [50]. 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 a mix of carbon-free (e.g., solar, wind) and carbon-based (e.g., coal, oil) energy sources (see Fig. 6.1). 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 [42, 78]. 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 [35, 174]. This carbon-aware scheduling paradigm is already being practiced at scale by companies such as Google [136, 83].

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

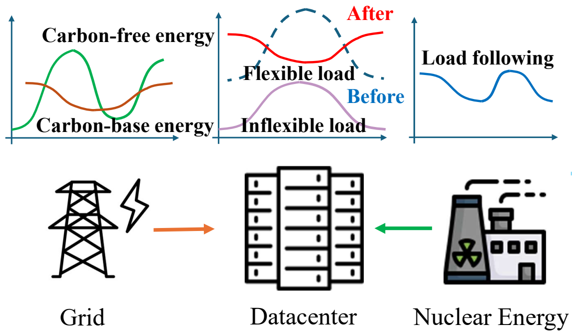


Figure 6.1: The framework of DCSMR

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 [146]. 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%.

6.2 Background and Related Work

Background on Nuclear Energy Generation: Fig. 6.2 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

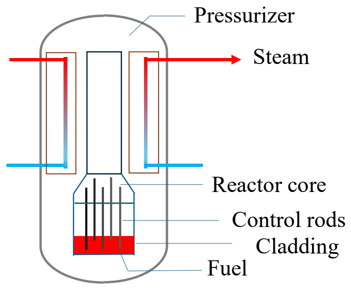


Figure 6.2: An LF-SMR

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 [97, 115]. 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) [152]. To address this, SMRs allow sufficient time for operators to implement reactivity changes to compensate for the xenon transient [129, 132].

Related Work: Taking SMRs as energy sources, recent studies have developed new models and optimizations on the electricity market [139], electricity-hydrogen integrated energy system [138] [92], electricity-heating multi-energy systems [140], 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 [151] 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 [33, 83]. There are studies on runtime optimization or day-ahead optimization. Runtime optimization minimizes the carbon by workload assignment [52, 66], frequency scaling [82], power capping [134], geographical server relocation [113], load shifting [103, 33], 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 [136, 83]. 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.

6.3 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 6.3.1. The datacenter power supply model is described in Section 6.3.2. 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 [45, 169]. 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 6.3.3 outlines the datacenter's workload-driven power demand. The various power-related costs are presented in Section 6.3.4. Building upon these models, we formulate the Day-ahead Datacenter Workload Planning with LF-SMR Problem in Section 6.3.5.

6.3.1 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 [132]. 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:

$$Rd_t + Up_t + St_t = 1, \quad \forall t, \tag{6.1}$$

$$Rd_t, Up_t, St_t \in \{0, 1\}, \quad \forall t. \tag{6.2}$$

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

[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.

(1) Ramping limits. The ramp-down and ramp-up limits of power output for an LF-SMR should satisfy Eqs. (6.3)-(6.4):

$$p_{t-1} - p_t \leqslant \overline{P}_{RD} \times Rd_t - \delta \times Up_t, \quad \forall t, \tag{6.3}$$

$$p_t - p_{t-1} \leqslant \overline{P}_{RU} \times Up_t - \delta \times Rd_t, \quad \forall t, \tag{6.4}$$

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. (6.5):

$$(Up_t - Up_{t-1}) \times T_h \leqslant \sum\nolimits_{tt=t-T_h}^{t-1} (St_t + Up_t), \quad \forall t, \tag{6.5}$$

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:

$$P_{MIN} \leqslant p_t \leqslant P_{MAX}, \quad \forall t, \tag{6.6}$$

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

6.3.2 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 [78]. [2] This stochastic nature is captured by modeling the actual carbon-free energy received, w_t, as:

$$w_t = (1 + \epsilon_t)W_t, \tag{6.7}$$

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:

$$g_{grid,t} = q_t + w_t. \tag{6.8}$$

6.3.3 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 [83]. Let jobs class be $c \in \mathcal{C}$. $s_{c,k}$ is the aggregate load of class c jobs submitted at time k. Then the total workload is:

$$z_t = \sum\nolimits_{c\in\mathcal{C}} \sum\nolimits_{k\in\mathcal{H}} Y_{k,c,t} \cdot s_{k,c}, \tag{6.9}$$

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

Each class $c \in \mathcal{C}$ 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 $h_c = 0$.

[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.

Then the workload planning problem is formulated as:

$$\sum\nolimits_{t\in\mathcal{T}} Y_{k,c,t} \geq 1, \quad \forall k \in \mathcal{H}, c \in \mathcal{C}, \tag{6.10}$$

$$Y_{k,c,t} = 0, \quad \forall k \in \mathcal{H}, c \in \mathcal{C}, t \notin \mathbb{Z}_{[k:k+h_c]}. \tag{6.11}$$

Eq. (6.10) ensures the full allocation of compute loads, while Eq. (6.11) 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:

$$v_t = e(z_t), \tag{6.12}$$

where $e(\cdot)$ represents the workload-to-energy consumption model, as described in [136] and [105].

6.3.4 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:

$$EPC = \sum\nolimits_{t=1}^{T} \alpha q_t + p_{co_2} q_t I_t + \kappa w_t, \tag{6.13}$$

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:

$$EGC = \sum_{t=1}^{T} \gamma p_t, \tag{6.14}$$

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 [135], e.g., by storage, etc., and this brings about a penalty cost.

$$EEC = \sum_{t=1}^{T} \beta(p_t - d_t), \tag{6.15}$$

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

6.3.5 Problem Formulation

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

$$\min EPC + EGC + EEC \tag{6.16}$$

$$\text{s.t.} \qquad (6.1) - (6.11), \tag{6.17}$$

$$p_t \geqslant d_t, \quad \forall t, \tag{6.18}$$

$$g_{grid,t} + d_t = v_t, \quad \forall t, \tag{6.19}$$

$$w_t + d_t \geqslant \eta v_t, \quad \forall t, \tag{6.20}$$

where η indicates the green factor, which lies within the range $[0, 1]$. Eq. (6.18) indicates that the consumption of nuclear power must exceed its generation. Eq. (6.19) specifies that the power supply should be equal to the demand. Eq. (6.20) 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.

6.4 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) [102] 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 [47]. 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.

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:

$$\mathbb{P}\left(Y_{w_{t+1}} \in \hat{C}(X_{w_{t+1}})\right) \geq \epsilon, \forall t, \tag{6.21}$$

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.

Algorithm 8 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.

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

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. (6.21).

3: Solve problem (6.16) - (6.20) , get the optimal SMR generation plan p_t and datacenter workload scheduling plan v_t. ▷ Stage 2: Optimization of DCSMR Problem

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

Corollary 6.4.1 (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. (6.20) 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. 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.

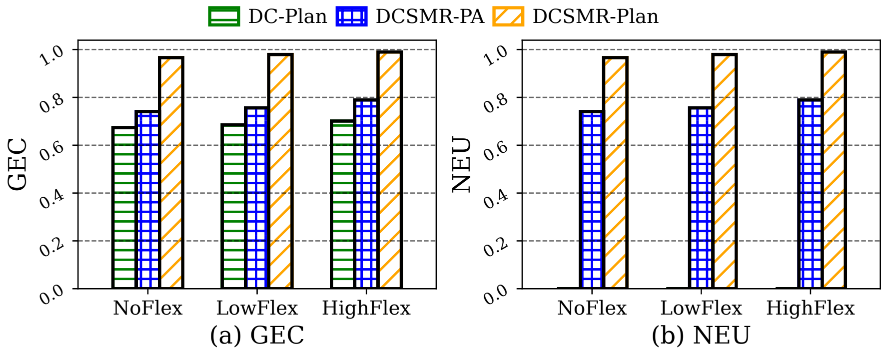


Figure 6.3: Performance under various workload flexibility

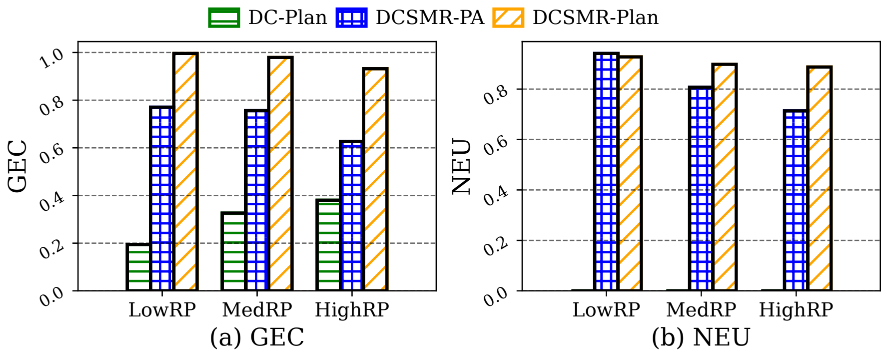


Figure 6.4: Performance under various renewable penetration

6.5 Evaluation

6.5.1 Evaluation Setup and Methodology

The datacenter workloads. We evaluate our model and algorithm using Google datacenter trace [146]. 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

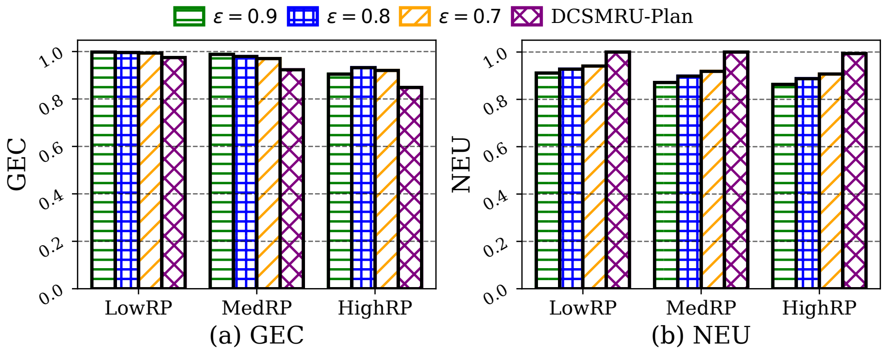


Figure 6.5: Performance of DCSMR-Plan under different confidence levels ϵ

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 factor of 20 MW. The ramp rate is set at 10%, and the hold time is specified as 3 hours, a standard practice widely adopted in nuclear power plant control [93] [141].

The power supply from the power grid. It is assumed that 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 [153]. 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) [33].* Proportion of renewable energy in total datacenter consumption. *(2) Nuclear Energy Utilization (NEU).* Proportion of nuclear energy consumed to total nuclear energy produced.

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 [184].

6.5.2 Evaluation results

Performance under different DC workload flexibility

Fig. 6.3 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.

Performance under different renewable penetration

Figure 6.4 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.

Impacts of uncertainty

Fig. 6.5 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, a trade-off must be made between ensuring green energy coverage and optimizing nuclear energy utilization.

6.6 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 [111]. 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.

Chapter 7

Conclusions and Suggestions for Future Research

In this chapter, we first conclude this thesis with a summarization of our original contributions in Section 7.1. Then we discuss some possible directions of future research in Section 7.2

7.1 Conclusions

This thesis presents optimization-based strategies aimed at decarbonizing energy-intensive sectors, specifically building integrated energy systems, building thermostatically controlled loads, and data center energy systems. Across these diverse applications, a central theme is the development of robust and data-driven frameworks designed to enhance operational efficiency, reduce carbon emissions towards climate neutrality goals, and effectively manage various sources of uncertainty inherent in these systems. By devising novel methodologies leveraging techniques such as robust learning, distributionally robust chance constraints, and conformal prediction, the research addresses limitations of traditional methods and provides crucial guarantees like statistical feasibility. Validated through comprehensive numerical studies using both synthetic and real-world data, the proposed approaches consistently demonstrate superior performance compared to existing literature, achieving significant reductions in energy costs, carbon emissions, and regret, while also improving control over constraint violations.

First, we have devised an et-zero optimization strategy to enable the efficient operation of diverse components and processes within building integrated energy systems (BIES), such as combined heat and power, energy storage, carbon capture, and the import of electricity from the primary grid. The proposed model aims to minimize the overall energy expenses while simultaneously curbing carbon emissions toward net-zero. Furthermore, to address the uncertainties inherent in the BIES, we have developed a robust learning-based optimization framework that provides statistical feasibility guarantees. The approach effectively handles the uncertainties and ensures reliable and robust operation by designing the appropriate uncertainty set. The proposed approach has been validated through numerical studies conducted using both synthetic and real data. Using synthetic data, our approach demonstrates improved control over the violation rate and leads to lower costs. Meanwhile, when applied to real-world data, the proposed approach achieves a significant reduction in regret percentage of over 8% for BIES control, compared to other methods in the literature. The algorithms in this work use an ellipsoidal uncertainty set to learn the shape of uncertainties. Future research could investigate the application of deep learning techniques to learning a more complex shape of the uncertainty set, thereby enhancing the precision. This chapter demonstrates the effectiveness of a robust learning-based approach with statistical feasibility guarantees for carbon-conscious BIES operation under uncertainty.

Second, we present a carbon-aware optimization strategy designed to enhance the efficient control of building TCLs. To mitigate uncertainties arising from outdoor temperature and base load forecast errors, we first formulate the TCL scheduling problem as a data-driven DRCC model under Wasserstein ambiguity sets, and then develop a novel reformulation strategy to recast the DRCC into a bilevel optimization structure. The proposed approach has been validated through simulation studies employing real-world data. Results have shown that the proposed approach yields a substantial reduction in regret percentage, surpassing 4.92% compared to the traditional approximation methods prevalent in existing literature for TCL control. Simulation results also demonstrate that the implementation of the carbon-aware mechanism can result in a noteworthy reduction in carbon emissions, 7.55% for our application. Furthermore, due to the alleviated conservativeness of the proposed approach, we show that significant cost and emission reductions can be achieved under any given comfortable range. This chapter highlights the benefits of a data-driven distributionally robust chance-constrained approach with a bilevel reformulation for carbon-aware and uncertainty-resilient TCLs.

Third, we study a carbon-aware optimization strategy that enables the efficient workload scheduling of data centers. The proposed model is designed to minimize overall energy expenses while simultaneously reducing carbon-based energy consumption toward 24/7 CEF. To address the uncertainties inherent in renewable generation predictions, a conformal prediction-based optimization framework is developed, providing statistical feasibility guarantees. The approach effectively manages uncertainties and ensures reliable and robust operation by integrating covariate information from the prediction model. The proposed framework is validated through numerical studies, demonstrating improved control over violation rates and achieving lower costs. Furthermore, the proposed approach achieves a cost reduction of over 6% in data center workload scheduling compared to traditional methods. This chapter presents a successful application of a conformal prediction-based optimization framework with statistical feasibility guarantees for carbon-conscious data center workload scheduling under uncertain renewable generation.

Finally, this thesis presents a novel day-ahead scheduling framework designed for co-optimizing data center workloads and load-following small modular reactor operations. This framework explicitly accounts for both nuclear operational constraints and the inherent uncertainty in renewable energy generation. Through extensive evaluation using real-world data center traces and simulated energy scenarios, it consistently demonstrates superior performance compared to baseline methods across varying levels of workload flexibility and renewable penetration. Key findings highlight that flexible workload scheduling, coordinated with LF-SMR control, can significantly enhance green energy coverage and nuclear energy utilization. This chapter shows that the proposed two-stage conformal prediction-based optimization algorithm maintains robustness even under high variability in renewable supply, offering a practical and scalable pathway for data centers to achieve demanding 24/7 carbon-free energy goals by effectively leveraging on-site nuclear power.

7.2 Future Research

The research presented in this thesis has successfully developed and validated data-driven optimization frameworks that provide statistical feasibility guarantees for decarbonization efforts in BIES, TCLs, and DCES under various uncertainties. Building upon this foundation, a key direction for future work lies in further enhancing the sophistication of uncertainty modeling and learning. While Chapter 3 explored learning ellipsoidal uncertainty sets, future research could leverage advanced machine learning and deep learning techniques to learn more complex, potentially non-convex or multimodal uncertainty set shapes directly from available data. This would offer a more accurate representation of uncertainty, leading to less conservative and potentially more effective optimization outcomes. Furthermore, extending the concept of integrating covariate information, as demonstrated in Chapter 5 with conformal prediction, to a broader range of energy systems is crucial. This would enable the development of truly contextual uncertainty models that dynamically adapt to real-time environmental, operational, or grid conditions. This is particularly relevant when considering uncertainty sources from sophisticated prediction models, including large-scale foundation models, whose prediction errors may exhibit complex patterns that vary significantly with context and input data, necessitating refined data-driven approaches for their characterization and integration into optimization.

Secondly, extending the application scope and computational efficiency of these advanced uncertainty-aware optimization methods is essential for addressing increasingly complex real-world energy systems. The methodologies developed in this thesis can be applied to and validated on more complex energy system architectures, such as multi-energy data centers that integrate diverse energy sources and conversion technologies. Applying these frameworks to larger-scale systems, such as optimizing operations across multiple buildings, a network of data centers, city-wide energy optimization problems, and vehicle-to-grid systems, also presents a significant opportunity. Successfully scaling these sophisticated formulations, especially those involving complex reformulations like the bilevel structure for DRCC problems, will require dedicated research into developing more computationally efficient algorithms, decomposition techniques, or potentially exploring machine learning-accelerated optimization approaches to handle the increased dimensionality and complexity inherent in these next-generation energy systems.

Third, a significant avenue for future research is the development of power-aware scheduling frameworks for AI-centric data centers. The proliferation of large language models (LLMs) is driving unprecedented investment in AI-centric data centers, creating a critical challenge in energy consumption. And different AI data center workload, such as LLM training, LLM fine-tuning, and LLM inference, have distinct power usage patterns. Collaboratively scaling GPU quantity and frequency can modulate power consumption without degrading performance, which provides load flexibility. Then it could be investigated how this capability can be synergistically paired with carbon-free power sources like nuclear energy. All in all, more work is required to realize efficient and reliable decarbonization strategies for future energy systems.

References

[1] `https://github.com/itsmeyijie/TCL-control`.

[2] Local businesses capture emissions and save energy through FortisBC pilot program, 2017.

[3] Scope 1 and scope 2 inventory guidance, 2022.

[4] Scope 3 inventory guidance, 2022.

[5] 24×7 carbon-free energy (part one): expectations and realities, 2023.

[6] California Independent System Operator, 2023.

[7] Carbon capture, 2023.

[8] Carbon capture reduces greenhouse gas emissions and saves money for Surrey school, 2023.

[9] CCS in Europe regional overview, 2023.

[10] Co2 emissions in 2023, 2023.

[11] Electricity Maps, 2023.

[12] New Group Trading Hourly Power Credits Forms in Bid to Clean Up the Grid, 2023.

[13] NYC skyscrapers turning to carbon capture to lessen climate change, 2023.

[14] The US Section 45Q Tax Credit for Carbon Oxide Sequestration: An Update , 2023.

[15] WattTime, 2023.

[16] What is The 45Q Tax Credit?, 2023.

[17] 24/7 carbon-free energy compact, 2024.

[18] Carbon Capture Pilot Program, 2024.

[19] Carbon capture technology for sustainable communities, 2024.

[20] CarbonQuest, 2024.

[21] Global energy and co2 status report 2018, 2024.

[22] NYC Multifamily Developer Taps CarbonQuest to Reduce Emissions, 2024.

[23] What is net zero?, 2024.

[24] Where do emissions come from? 4 charts explain greenhouse gas emissions by sector, 2024.

[25] Carbon awareness, 2025.

[26] Co2 capture and utilisation, 2025.

[27] Grid carbon intensity, 2025.

[28] Net-zero carbon, 2025.

[29] Record-breaking annual growth in renewable power capacity, 2025.

[30] Soletair power carbon capture technology, 2025.

[31] What is carbon capture and storage?, 2025.

[32] Timothy Gardner. Meta seeks nuclear power developers for reactors to start in early 2030s, 2024.

[33] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 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, pages 118–132, Vancouver BC Canada, March 2023.

[34] Rajendra Adhikari, Manisa Pipattanasomporn, and Saifur Rahman. Heuristic algorithms for aggregated hvac control via smart thermostats for regulation service. *IEEE Transactions on Smart Grid*, 11(3):2023–2032, May 2019.

[35] Mehdi Ahmadi, Lukas Knorr, and Henning Meschede. Improvement of wind power utilization through flexible operation of data center in wind parks. *Renewable Energy*, page 123073, 2025.

[36] Shabbir Ahmed, James Luedtke, Yongjia Song, and Weijun Xie. Nonanticipative duality, relaxations, and formulations for chance-constrained stochastic programs. *Mathematical Programming*, 162(1):51–81, May 2017.

[37] Joseph Aldy, William Pizer, Massimo Tavoni, Lara Aleluia Reis, Keigo Akimoto, Geoffrey Blanford, Carlo Carraro, Leon E Clarke, James Edmonds, Gokul C Iyer, et al. Economic tools to promote transparency and comparability in the paris agreement. *Nature Climate Change*, 6(11):1000–1004, 2016.

[38] Saeed Alhadhrami, Gabriel J Soto, and Ben Lindley. Dispatch analysis of flexible power operation with multi-unit small modular reactors. *Energy*, 280:128107, 2023.

[39] Lazaro Alvarado-Barrios, Alvaro Rodríguez del Nozal, Juan Boza Valerino, Ignacio García Vera, and Jose L Martínez-Ramos. Stochastic unit commitment in

microgrids: Influence of the load forecasting error and the availability of energy storage. *Renewable Energy*, 146:2060–2069, 2020.

[40] Abigail M Andrews and Rishee K Jain. A policy landscape analysis of demand flexibility driven building decarbonization: a case study of new york city, usa. *Energy*, 2004:2965, 2022.

[41] Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

[42] José Arellano and Miguel Carrión. Electricity procurement of large consumers considering power-purchase agreements. *Energy Reports*, 9:5384–5396, 2023.

[43] Toshi H Arimura and Tatsuya Abe. The impact of the tokyo emissions trading scheme on office buildings: what factor contributed to the emission reduction? *Environmental Economics and Policy Studies*, 23:517–533, 2021.

[44] Adriano Arrigo, Christos Ordoudis, Jalal Kazempour, Zacharie De Grève, Jean-François Toubeau, and François Vallée. Wasserstein distributionally robust chance-constrained optimization for energy and reserve dispatch: An exact and physically-bounded formulation. *European Journal of Operational Research*, 296(1):304–322, January 2022.

[45] Cigdem Aslan and Tim Irwin. 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.

[46] Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

[47] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

[48] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167:235–292, 2018.

[49] Dimitris Bertsimas, Thodoris Koukouvinos, and Angelos Georgios Koulouras. Constructing uncertainty sets from covariates in power systems. *IEEE Transactions on Power Systems*, 2025.

[50] Lori Bird, Debra Lew, Michael Milligan, E Maria Carlini, Ana Estanqueiro, Damian Flynn, Emilio Gomez-Lazaro, Hannele Holttinen, Nickie Menemenlis, Antje Orths, et al. Wind and solar energy curtailment: A review of international experience. *Renewable and Sustainable Energy Reviews*, 65:577–586, 2016.

[51] Rafael Blanquero, Emilio Carrizosa, and Nuria Gómez-Vargas. Contextual uncertainty sets in robust linear optimization. 2023.

[52] Roozbeh Bostandoost, Walid A Hanafy, Adam Lechowicz, Noman Bashir, Prashant Shenoy, and Mohammad Hajiesmaili. Data-driven algorithm selection for carbon-aware scheduling. *ACM SIGENERGY Energy Informatics Review*, 4(5):148–153, 2024.

[53] Sibylle Braungardt, Veit Bürger, and Benjamin Köhler. Carbon pricing and complementary policies—consistency of the policy mix for decarbonizing buildings in germany. *Energies*, 14(21):7143, 2021.

[54] Kenneth Bruninx, Yury Dvorkin, Erik Delarue, William D'haeseleer, and Daniel S Kirschen. Valuing demand response controllability via chance constrained programming. *IEEE Transactions on Sustainable Energy*, 9(1):178–187, January 2017.

[55] Giuseppe Calafiore and Marco C Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102:25–46, 2005.

[56] Giuseppe Carlo Calafiore and Marco C Campi. The scenario approach to robust control design. *IEEE Transactions on automatic control*, 51(5):742–753, May 2006.

[57] Giuseppe Carlo Calafiore and L El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(4):1–22, December 2006.

[58] Ge Chen, Hongcai Zhang, Hongxun Hui, and Yonghua Song. Fast wasserstein-distance-based distributionally robust chance-constrained power dispatch for multi-zone hvac systems. *IEEE Transactions on Smart Grid*, 12(5):4016–4028, 2021.

[59] Xin Chen. Enhance low-carbon power system operation via carbon-aware demand response. *Energy Internet*, 1(2):141–149, 2024.

[60] Xin Chen and Na Li. Leveraging two-stage adaptive robust optimization for power flexibility aggregation. *IEEE Transactions on Smart Grid*, 12(5):3954–3965, September 2021.

[61] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls. *Operations Research*, 72(1):410–424, July 2022.

[62] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. On approximations of data-driven chance constrained programs over wasserstein balls. *Operations Research Letters*, 51(3):226–233, May 2023.

[63] Zhi Chen, Lei Wu, and Yong Fu. Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization. *IEEE transactions on smart grid*, 3(4):1822–1831, December 2012.

[64] Zhi Chen, Lei Wu, and Zuyi Li. Electric demand response management for distributed large-scale internet data centers. *IEEE transactions on Smart Grid*, 5(2):651–661, 2013.

[65] Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. Data-driven conditional robust optimization. *Advances in Neural Information Processing Systems*, 35:9525–9537, 2022.

[66] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. Reducing the carbon impact of generative ai inference (today and in 2035). In *Proceedings of the 2nd workshop on sustainable computer systems*, pages 1–7, 2023.

[67] Okan Ciftci, Mahdi Mehrtash, and Amin Kargarian. Data-driven nonparametric chance-constrained optimization for microgrid energy management. *IEEE Transactions on Industrial Informatics*, 16(4):2447–2457, 2019.

[68] Austin R. Coffman, Neil Cammardella, Prabir Barooah, and Sean Meyn. Aggregate flexibility capacity of tcls with cycling constraints. *IEEE Transactions on Power Systems*, 38(1):52–62, 2023.

[69] Samuel Córdova, Claudio A Cañizares, Álvaro Lorca, and Daniel E Olivares. Aggregate modeling of thermostatically controlled loads for microgrid energy management systems. *IEEE Transactions on Smart Grid*, 14(6):4169–4181, 2023.

[70] Avijit Das and Zhen Ni. A computationally efficient optimization approach for battery systems in islanded microgrid. *IEEE Transactions on Smart Grid*, 9(6):6489–6499, 2017.

[71] YF Du, Lin Jiang, Chao Duan, YZ Li, and JS Smith. Energy consumption scheduling of hvac considering weather forecast error through the distributionally robust approach. *IEEE Transactions on Industrial Informatics*, 14(3):846–857, March 2018.

[72] Pengfei Duan, Bingxu Zhao, Xinghui Zhang, and Mengdan Fen. A day-ahead optimal operation strategy for integrated energy systems in multi-public buildings based on cooperative game. *Energy*, 275(14):127395, July 2023.

[73] Samy Faddel. Modeling and robust optimization of commercial air-water hvac system. *IEEE Transactions on Industry Applications*, 2023.

[74] Mohammad Hassan Fathollahzadeh and Paulo Cesar Tabares-Velasco. Integrated framework for optimization of air-and water-side hvac systems to minimize electric utility cost of existing commercial districts. *Energy and Buildings*, 273(20):112328, October 2022.

[75] Najmeh Forouzandehmehr, Samir M Perlaza, Zhu Han, and H Vincent Poor. A satisfaction game for heating, ventilation and air conditioning control of smart buildings. In *2013 IEEE Global Communications Conference (GLOBECOM)*, pages 3164–3169. IEEE, 2013.

[76] Robin Gaster. Small modular reactors: A realist approach to the future of nuclear power. Technical report, Information Technology and Innovation Foundation, 2025.

[77] Mahdi Ghamkhari and Hamed Mohsenian-Rad. Energy and performance management of green data centers: A profit maximization approach. *IEEE transactions on Smart Grid*, 4(2):1017–1025, 2013.

[78] Yashar Ghiassi-Farrokhfal, Wolfgang Ketter, and John Collins. Making green power purchase agreements more predictable and reliable for companies. *Decision Support Systems*, 144:113514, 2021.

[79] Goldman Sachs Research . AI to drive 165% increase in data center power demand by 2030, 2025.

[80] JM González-Sopeña, V Pakrashi, and B Ghosh. An overview of performance evaluation metrics for short-term statistical wind power forecasting. *Renewable and Sustainable Energy Reviews*, 138, 2021.

[81] Gordon Noble, Fiona Berry. Power-hungry AI is driving a surge in tech giant carbon emissions. Nobody knows what to do about it, 2024.

[82] Diandian Gu, Yihao Zhao, Peng Sun, Xin Jin, and Xuanzhe Liu. Greenflow: A carbon-efficient scheduler for deep learning workloads. *IEEE Transactions on Parallel and Distributed Systems*, 2024.

[83] Sophie Hall, Francesco Micheli, Giuseppe Belgioioso, Ana Radovanović, and Florian Dörfler. Carbon-aware computing for data centers with probabilistic performance guarantees. *arXiv preprint arXiv:2410.21510*, 2024.

[84] Savannah Goodman Hallie Cramer. Accelerating a carbon-free future with hourly energy tracking, 2024.

[85] Miguel Heleno, Manuel A Matos, and JA Pecas Lopes. Availability and flexibility of loads for the provision of reserve. *IEEE Transactions on Smart Grid*, 6(2):667–674, March 2014.

[86] Helmuth Boeck. Why nuclear power is safer than ever, 2022.

[87] Helmuth Boeck. Safety of nuclear power reactors, 2025.

[88] Lars Herre, Behrouz Nourozi, Mohammad Reza Hesamzadeh, Qian Wang, and Lennart Söder. Provision of multiple services with controllable loads as multi-area thermal energy storage. *Journal of Energy Storage*, 63:107062, 2023.

[89] L Jeff Hong, Zhiyuan Huang, and Henry Lam. Learning-based robust optimization: Procedures and statistical guarantees. *Management Science*, 67(6):3447–3467, December 2021.

[90] Bowen Huang, Sen Huang, Xu Ma, Srinivas Katipamula, Di Wu, and Robert Lutes. Stochastic scheduling for commercial building cooling systems: considering uncertainty in zone temperature prediction. *Applied Energy*, 346:121367, 2023.

[91] Hongxu Huang, Zhengmao Li, LP Mohasha Isuru Sampath, Jiawei Yang, Hung D Nguyen, Hoay Beng Gooi, Rui Liang, and Dunwei Gong. Blockchain-enabled carbon and energy trading for network-constrained coal mines with uncertainties. *IEEE Transactions on Sustainable Energy*, 14(3):634 – 1647, 2023.

[92] Roshni Anna Jacob and Jie Zhang. 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.

[93] Jesse D Jenkins, Zhi Zhou, Roberto Ponciroli, Richard B Vilim, Francesco Ganda, Fernando de Sisternes, and Audun Botterud. The benefits of nuclear flexibility in power system operations with renewable energy. *Applied energy*, 222:872–884, 2018.

[94] Nan Jiang and Weijun Xie. Also-x and also-x+: Better convex approximations for chance constrained programs. *Operations Research*, 70(6):3581–3600, Feburary 2022.

[95] Wenqian Jiang, Chenbei Lu, and Chenye Wu. Robust scheduling of thermostatically controlled loads with statistically feasible guarantees. *IEEE Transactions on Smart Grid*, 14(5):3561–3572, 2023.

[96] Joao da Silva . Google turns to nuclear to power AI data centres, 2024.

[97] Jan Horst Keppler and Marco Cometto. Nuclear energy and renewables: system effects in low-carbon electricity systems. Technical report, 2012.

[98] Abbas Kiani and Nirwan Ansari. Profit maximization for geographically dispersed green data centers. *IEEE Transactions on Smart Grid*, 9(2):703–711, 2016.

[99] Rudolf Kruse, Sanaz Mostaghim, Christian Borgelt, Christian Braune, and Matthias Steinbrecher. Multi-layer perceptrons. In *Computational intelligence: a methodological introduction*, pages 53–124. Springer, 2022.

[100] Soongeol Kwon, Lewis Ntaimo, and Natarajan Gautam. Demand response in data centers: Integration of server provisioning and power procurement. *IEEE transactions on Smart Grid*, 10(5):4928–4938, 2018.

[101] Laure de Roucy-Rochegonde and Adrien Buffard. Ai, data centers and energy demand: Reassessing and exploring the trends, 2025.

[102] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

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

[104] Jie Li and Wenbo Qi. Toward optimal operation of internet data center microgrid. *IEEE Transactions on Smart Grid*, 9(2):971–979, 2016.

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

[106] Xianyu Li, Fenglian Dong, Zhiwei Wei, and Chao Shang. Data-driven contextual robust optimization based on support vector clustering. *Computers & Chemical Engineering*, page 109004, 2025.

[107] Yanbin Li, Yanting Sun, Jiechao Liu, Chang Liu, and Feng Zhang. A data driven robust optimization model for scheduling near-zero carbon emission power plant considering the wind power output uncertainties and electricity-carbon market. *Energy*, 279:128053, 2023.

[108] Yanbin Li, Feng Zhang, Yun Li, and Yuwei Wang. An improved two-stage robust optimization model for cchp-p2g microgrid system considering multi-energy operation under wind power outputs uncertainties. *Energy*, 223(10):120048, May 2021.

[109] Yong Li, Yao Zou, Yi Tan, Yijia Cao, Xindong Liu, Mohammad Shahidehpour, Shiming Tian, and Fanpeng Bu. Optimal stochastic operation of integrated low-carbon electric power, natural gas, and heat delivery system. *IEEE Transactions on Sustainable Energy*, 9(1):273–283, 2017.

[110] Yuzhuo Li, Mariam Mughees, Yize Chen, and Yunwei Ryan Li. The unseen ai disruptions for power grids: Llm-induced transients. *arXiv preprint arXiv:2409.11416*, 2024.

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

[112] Liuzixuan Lin, Rajini Wijayawardana, Varsha Rao, Hai Nguyen, Emmanuel Wedan GNIBGA, and Andrew A Chien. 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*, pages 434–441, 2024.

[113] Yejia Liu, Pengfei Li, Daniel Wong, and Shaolei Ren. Geographical server relocation: Opportunities and challenges. *ACM SIGENERGY Energy Informatics Review*, 4(5):34–43, 2024.

[114] Don Rukmal Liyanage, Kasun Hewage, Hirushie Karunathilake, and Rehan Sadiq. Feasibility of carbon-capturing in building heating systems: A life cycle thinking-based approach. *International Journal of Greenhouse Gas Control*, 132:104056, 2024.

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

[116] Junchen Lu et al. Small modular reactors: A viable path to sustainable data centers in the age of artificial intelligence. *The Frontiers of Society, Science and Technology*, 6(11), 2024.

[117] James Luedtke. A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1):219–244, 2014.

[118] James Luedtke and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.

[119] Jianying Luo, Lei Rao, and Xue Liu. Temporal load balancing with service delay guarantees for data center energy cost optimization. *IEEE Transactions on Parallel and Distributed Systems*, 25(3):775–784, 2013.

[120] Diptyaroop Maji, Prashant Shenoy, and Ramesh K Sitaraman. Carboncast: multi-day forecasting of grid carbon intensity. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 198–207, Boston, Massachusetts, USA, November 2022.

[121] Diptyaroop Maji, Ramesh K Sitaraman, and Prashant Shenoy. Dacf: day-ahead carbon intensity forecasting of power grids using machine learning. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*, pages 188–192, Virtual Event, June 2022.

[122] MCE Community Choice Energy . MCE Light Green Non-Residential Rates, 2023.

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

[124] Ali Mobasseri, Marcos Tostado-Véliz, Ali Asghar Ghadimi, Mohammad Reza Miveh, and Francisco Jurado. Multi-energy microgrid optimal operation with integrated power to gas technology considering uncertainties. *Journal of Cleaner Production*, 333(4):130174, January 2022.

[125] Natalie Sherman. Microsoft chooses infamous nuclear site for AI power, 2024.

[126] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, October 2007.

[127] Anh Tuan Nguyen, Duy Hoang Pham, Bee Lan Oo, Mattheos Santamouris, Yonghan Ahn, and Benson TH Lim. Modelling building hvac control strategies using a deep reinforcement learning approach. *Energy and Buildings*, 310:114065, 2024.

[128] NVIDIA. NVIDIA DGX B200,The foundation for your AI factory., 2024.

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

[130] Yash P Patel, Sahana Rayan, and Ambuj Tewari. Conformal contextual robust optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2485–2493. PMLR, 2024.

[131] Sarin Raj Pokhrel, Kasun Hewage, Gyan Chhipi-Shrestha, Hirushie Karunathilake, Eric Li, and Rehan Sadiq. Carbon capturing for emissions reduction at building level: A market assessment from a building management perspective. *Journal of Cleaner Production*, 294:126323, 2021.

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

[133] Álvaro Porras, Concepción Domínguez, Juan Miguel Morales, and Salvador Pineda. Tight and compact sample average approximation for joint chance-constrained problems with applications to optimal power flow. *INFORMS Journal on Computing*, 35(6):1454–1469, 2023.

[134] Haoran Qiu, Linghao Zhang, Chen Wang2 Hubertus Franke, Zbigniew T Kalbarczyk, and Ravishankar K Iyer. 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)*, 2023.

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

[136] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care,

et al. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems*, 38(2):1270–1280, 2022.

[137] Hamed Rahimian and Bernardo Pagnoncelli. Contextual chance-constrained programming. *Available on Optimization Online*, 2020.

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

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

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

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

[142] Rockafellar, Uryasev, Nemirovsky, and Shapiro. Chance constrained optimization, 2023.

[143] Guangchun Ruan, Haiwang Zhong, Guanglun Zhang, Yiliu He, Xuan Wang, and Tianjiao Pu. Review of learning-assisted power system optimization. *CSEE Journal of Power and Energy Systems*, 7(2):221–231, 2020.

[144] Hossein Saberi, Cuo Zhang, and Zhao Yang Dong. Data-driven distributionally robust hierarchical coordination for home energy management. *IEEE Transactions on Smart Grid*, 12(5):4090–4101, September 2021.

[145] Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

[146] Varun Sakalkar, Vasileios Kontorinis, David Landhuis, Shaohong Li, Darren De Ronde, Thomas Blooming, Anand Ramesh, James Kennedy, Christopher Malone, Jimmy Clidaras, et al. 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*, pages 497–511, 2020.

[147] Fahad Ben Salamah. *Building-Integrated Carbon Capture: A Study on the Design and Potential of Carbon Emission Offsets from Buildings.* PhD thesis, Arizona State University, 2021.

[148] Chengcheng Shao, Chenjia Feng, Mohammad Shahidehpour, Quan Zhou, Xiuli Wang, and Xifan Wang. Optimal stochastic operation of integrated electric power and renewable energy with vehicle-based hydrogen energy system. *IEEE Transactions on Power Systems*, 36(5):4310–4321, September 2021.

[149] Sumedha Sharma, Ashu Verma, Yan Xu, and Bijaya Ketan Panigrahi. Robustly coordinated bi-level energy management of a multi-energy building under multiple uncertainties. *IEEE Transactions on Sustainable Energy*, 12(1):3–13, 2019.

[150] Xiangnan Song, Meng Shen, Yujie Lu, Liyin Shen, and Hongyang Zhang. How to effectively guide carbon reduction behavior of building owners under emission trading scheme? an evolutionary game-based study. *Environmental Impact Assessment Review*, 90:106624, 2021.

[151] Gabriel Jose Soto Gonzalez, Botros Naseif Hanna Bishara Hanna, Jakub Toman, Nahuel Guaita, Paul W Talbot, Christopher Shawn Lohse, and Aaron S

Epiney. 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), 2024.

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

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

[154] Su Su, Zening Li, Xiaolong Jin, Koji Yamashita, Mingchao Xia, and Qifang Chen. Energy management for active distribution network incorporating office buildings based on chance-constrained programming. *International Journal of Electrical Power & Energy Systems*, 134(1):107360, January 2022.

[155] Chunlin Sun, Linyu Liu, and Xiaocheng Li. Predict-then-calibrate: A new perspective of robust contextual lp. *Advances in Neural Information Processing Systems*, 36:17713–17741, 2023.

[156] Gaiping Sun, Chuanwen Jiang, Xu Wang, and Xiu Yang. Short-term building load forecast based on a data-mining feature selection and lstm-rnn method. *IEEJ Transactions on Electrical and Electronic Engineering*, 15(7):1002–1010, July 2020.

[157] Vahid Sohrabi Tabar, Mehrdad Tarafdar Hagh, and Mehdi Ahmadi Jirdehi. Achieving a nearly zero energy structure by a novel framework including energy recovery and conversion, carbon capture and demand response. *Energy and Buildings*, 230:110563, 2021.

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

[159] Pham Dinh Tao and LT Hoai An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997.

[160] Anders Thavlov and Henrik W Bindner. Utilization of flexible demand in a virtual power plant set-up. *IEEE Transactions on Smart Grid*, 6(2):640–647, March 2014.

[161] The National Renewable Energy Laboratory . End Use Load Profiles for the U.S. Building Stock.

[162] The United Nations. 24/7 carbon-free energy compact, 2025.

[163] The White House. FACT SHEET: President Biden Sets 2030 Greenhouse Gas Pollution Reduction Target Aimed at Creating Good-Paying Union Jobs and Securing U.S. Leadership on Clean Energy Technologies, 2021.

[164] Guanyu Tian and Qun Zhou Sun. Chance constrained distributionally robust optimal hvac scheduling for commercial building demand response. In *2022 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, Singapore, April 2022. IEEE.

[165] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the next generation. In *Proceedings of the fifteenth European conference on computer systems*, pages 1–14, Heraklion, Greece, April 2020.

[166] Hashem Amini Toosi, Monica Lavagna, Fabrizio Leonforte, Claudio Del Pero, and Niccolò Aste. Building decarbonization: Assessing the potential of building-integrated photovoltaics and thermal energy storage systems. *Energy Reports*, 8(16):574–581, December 2022.

[167] U.S. Environmental Protection Agency. Climate Change Indicators: Greenhouse Gases.

[168] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

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

[170] Cheng Wang, Ning Gao, Jianhui Wang, Nan Jia, Tianshu Bi, and Kenneth Martin. Robust operation of a water-energy nexus: A multi-energy perspective. *IEEE Transactions on Sustainable Energy*, 11(4):2698–2712, 2020.

[171] Jiaqiang Wang, Zhenlin Huang, Chang Yue, Quan Zhang, and Peng Wang. Various uncertainties self-correction method for the supervisory control of a hybrid cooling system in data centers. *Journal of Building Engineering*, 42(10):102830, October 2021.

[172] Xiaowei Wang and Yue Chen. Carbon-aware quantification of real-time aggregate power flexibility of electric vehicles. *arXiv preprint arXiv:2409.05597*, 2024.

[173] Yingjie Wang, Qi Zeng, Zhongbei Tian, Yuefang Du, Pingliang Zeng, and Lin Jiang. Distributionally robust energy consumption scheduling of hvac considering the uncertainty of outdoor temperature and human activities. *CSEE Journal of Power and Energy Systems*, 9(3):896–909, May 2022.

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

[175] Junghoon Woo, Ridah Fatima, Charles J Kibert, Richard E Newman, Yifeng Tian, and Ravi S Srinivasan. Applying blockchain technology for building energy performance measurement, reporting, and verification (mrv) and the car-

bon credit market: A review of the literature. *Building and Environment*, 205:108199, 2021.

[176] Qunli Wu and Chunxiang Li. Modeling and operation optimization of hydrogen-based integrated energy system with refined power-to-gas and carbon-capture-storage technologies under carbon trading. *Energy*, 270(9):126832, May 2023.

[177] Zhi Wu, Suyang Zhou, Jianing Li, and Xiao-Ping Zhang. Real-time scheduling of residential appliances via conditional risk-at-value. *IEEE Transactions on Smart Grid*, 5(3):1282–1291, May 2014.

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

[179] Jiali Xing, Bilge Acun, Aditya Sundarrajan, David Brooks, Manoj Chakkaravarthy, Nikky Avila, Carole-Jean Wu, and Benjamin C Lee. Carbon responder: Coordinating demand response for the datacenter fleet. *arXiv preprint arXiv:2311.08589*, 2023.

[180] Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55076–55099, 2024.

[181] Mingyu Yan, Ning Zhang, Xiaomeng Ai, Mohammad Shahidehpour, Chongqing Kang, and Jinyu Wen. Robust two-stage regional-district scheduling of multi-carrier energy systems with a large penetration of wind power. *IEEE Transactions on Sustainable Energy*, 10(3):1227–1239, 2018.

[182] Jun Yang and Changqi Su. Robust optimization of microgrid based on renewable distributed power generation and load demand uncertainty. *Energy*, 223:120043, 2021.

[183] Lun Yang, Yinliang Xu, Wei Gu, and Hongbin Sun. Distributionally robust chance-constrained optimal power-gas flow under bidirectional interactions considering uncertain wind power. *IEEE Transactions on Smart Grid*, 12(2):1722–1735, March 2020.

[184] Yijie Yang. Github, dcsmr-plan, 2025.

[185] Yijie Yang, Jian Shi, Dan Wang, Chenye Wu, and Zhu Han. Net-zero scheduling of multi-energy building energy systems: A learning-based robust optimization approach with statistical guarantees. *IEEE Transactions on Sustainable Energy*, 2024.

[186] Yuqiang Yang, Lei Li, Weitao Tan, Hai Hu, Weiliang Zhang, and Zhenzhi Lin. A thermostatically controlled load cluster regulation method for peak-valley load balancing. In *International Conference on Frontiers of Energy and Environment Engineering*, pages 349–359. Springer, 2023.

[187] Zhihao Yang, Anupam Trivedi, Haoming Liu, Ming Ni, and Dipti Srinivasan. Two-stage robust optimization strategy for spatially-temporally correlated data centers with data-driven uncertainty sets. *Electric Power Systems Research*, 221:109443, 2023.

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

[189] Junyi Zhai, Sheng Wang, Lei Guo, Yuning Jiang, Zhongjian Kang, and Colin N Jones. Data-driven distributionally robust joint chance-constrained energy management for multi-energy microgrid. *Applied Energy*, 326:119939, 2022.

[190] Chen Zhang, Yiliang Xie, Hongxia Zhang, Yujie Gu, and Xiongwen Zhang. Optimal design and performance assessment for a solar powered electricity, heating and hydrogen integrated energy system. *Energy*, 262(1):125453, January 2023.

[191] Cuo Zhang, Yan Xu, Zhengmao Li, and Zhao Yang Dong. Robustly coordinated operation of a multi-energy microgrid with flexible electric and thermal loads. *IEEE Transactions on Smart Grid*, 10(3):2765–2775, 2018.

[192] Hui Zhang, Jiye Wang, Xiongwen Zhao, Jingqi Yang, et al. Modeling a hydrogen-based sustainable multi-carrier energy system using a multi-objective optimization considering embedded joint chance constraints. *Energy*, 278(17):127643, September 2023.

[193] Xiaoyang Zhang and Dan Wang. A gnn-based day ahead carbon intensity forecasting model for cross-border power grids. In *Proceedings of the 14th ACM International Conference on Future Energy Systems*, pages 361–373, 2023.

[194] Zhenhe Zhang, Jun Zheng, and Guchuan Zhu. Event-triggered power tracking control of heterogeneous tcl populations. *IEEE Transactions on Smart Grid*, 15(4), 2024.

[195] Yu Zheng and Bing Zhang. The impact of carbon market on city greening: Quasi-experimental evidence from china. *Resources, Conservation and Recycling*, 193:106960, 2023.

[196] Junjie Zhong, Yong Li, Yijia Cao, Yi Tan, Yanjian Peng, Yicheng Zhou, Yosuke Nakanishi, and Zhengmao Li. Robust coordinated optimization with adaptive uncertainty set for a multi-energy microgrid. *IEEE Transactions on Sustainable Energy*, 14(1):111–124, 2022.

[197] Yinyu Zhou and Bo Zeng. Bi-level planning method for internet data center and distribution network considering demand response. In *2022 IEEE 5th International Electrical and Energy Conference (CIEEC)*, pages 3270–3275. IEEE, 2022.