



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Unlocking Grid Flexibility of Distributed Energy Resources

Yihong Zhou



Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2025

Abstract

Power systems are increasingly integrating distributed energy resources (DERs), such as distributed renewable energy sources (RES), heat pumps (HPs), and data centres. While DERs can strain grid capacity, they can also offer power flexibility to support stable system operation. To unlock this opportunity, this thesis focuses on tackling unresolved challenges in realising DER flexibility.

From a technical perspective, fully unlocking the DER flexibility requires optimal scheduling. Although extensive work has studied DER scheduling, a key challenge remains in handling the inherent uncertainty of DERs, which arises from weather dependence and human behaviour and poses a challenge to reliable power system operation. Joint chance constraints (JCC) represent a promising approach by explicitly guaranteeing the probability of joint constraint satisfaction. This method has been widely applied both in standard power system scheduling and in specific frameworks designed for efficient large-scale scheduling, such as the hierarchical scheduling regime known as DER aggregation. However, existing integrations of JCC are computationally demanding, which limits their scalability. This challenge leads to the first research question: *Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering uncertainty?*

Beyond the technical challenge of scheduling DERs, existing work has investigated the technical, economic and social impacts of many DERs. These studies also represent a crucial dimension that supports the design of effective and efficient roadmaps to realise DER flexibility. However, two important types of DERs, namely HPs and data centres, remain under-investigated. Given the pivotal role that HPs and data centres are expected to play in future power systems, this knowledge gap poses a significant challenge in unlocking the full value of DER flexibility. In particular, the impact of HPs on fuel poverty remains insufficiently investigated. Fuel poverty refers to a status where households struggle to heat their homes to a comfortable standard, representing a critical economic and social issue. The approximately 300% running efficiency of HPs makes them an especially attractive alternative for households reliant on expensive off-gas heating sources, such as heating oil and electric resistive heating. These households are often at high risk of fuel poverty, further increasing the need for deploying lower-cost, efficient heating solutions such as HPs. However, the HP impact on fuel poverty remains unclear. Furthermore, the large-scale deployment of HPs may also threaten the distribution network, and this network impact, which is interlinked with the fuel poverty impact, has also been under-explored. These gaps motivate the second research question: *Q2: What are the fuel poverty and distribution network impacts of replacing domestic off-gas heating systems with HPs?*

For data centres, there is limited understanding in whether they can bring a new and significant source grid flexibility, particularly for the emerging data centres dedicated to Artificial Intelligence (AI). The recent expansion of the AI sector is driving the deployment of AI-focused data centres, which are more energy-intensive than traditional general-purpose data centres due to the heavy reliance on GPUs. However, it remains unclear whether AI-focused data centres represent a new and significant source of power system flexibility. Furthermore, the lack of quantification of the associated cost means that it is also uncertain whether data centre operators could be effectively incentivised to offer this flexibility. These gaps motivate the third research question: *Q3: Do AI-focused data centres have a different potential for power system flexibility compared to general-purpose data centres, and what is their cost of flexibility provision?*

To address Q1, this thesis develops novel methods for the efficient integration of JCC in both standard DER scheduling problems and DER aggregation. For standard DER scheduling problems, the focus is placed on a particular JCC structure: the Wasserstein distributionally robust JCC (WDRJCC) with right-hand-side (RHS) uncertainty. This formulation can be applied to ensure power system reserve sufficiency and to ensure network thermal limits with enhanced robustness. This thesis proposes a strengthened and faster linear approximation (SFLA) for the RHS-WDRJCC, demonstrating up to a 100× speedup over traditional methods, such as worst-case conditional value-at-risk (CVaR). For DER aggregation, this thesis develops a tractable method to integrate WDRJCC and classical JCC with a two-stage structure. The proposed method is shown to achieve a better balance between conservativeness and reliability compared to existing approaches. In addition, the computational complexity of the proposed method remains low and does not increase with the number of DERs nor the size of the network, thereby facilitating its applicability to real-world, industry-scale problems.

To tackle Q2, this thesis evaluates the fuel poverty and distribution network impact of HP replacing off-gas heating in the UK. A novel evaluation method is proposed to support this evaluation by integrating publicly available datasets from a wide range of sources, and the method can be extended to other countries. The results show that replacing off-gas heating with HPs could lift up to 20% of households out of fuel poverty in specific UK regions, providing both social and economic incentives for HP deployment. However, this replacement would incur significant distribution network upgrade costs in certain areas, underscoring the need to realise HP flexibility.

Finally, to address Q3, this thesis proposes a novel method to evaluate both the maximum flexibility and the corresponding cost for data centres providing power system services, ranging from short-duration, low-frequency services to long-duration, high-frequency services. To ensure robust conclusions, the proposed method is applied to datasets from 14 real-world data centres and three mainstream cloud computing platforms. The results demonstrate that

AI-focused, GPU-intensive data centres can deliver greater flexibility at up to 50% lower cost compared to traditional CPU-intensive data centres. This insight provides motivation for both power grid operators and AI data centre operators to collaborate in future schemes aimed at realising data centre flexibility.

Overall, this thesis contributes to the understanding of how to unlock DER flexibility in future power systems. The proposed methods for efficient JCC integration and the evaluation of HP and data centre impacts provide valuable insights for both researchers and practitioners. The findings highlight the potential of DERs to support reliable and efficient power system operation, while also addressing critical social issues such as fuel poverty.

Lay Summary

The future of our energy systems is being shaped by new technologies like renewable energy sources, heat pumps (HPs) for heating, and data centres powering our digital world. These technologies, known as distributed energy resources (DERs), are often connected to power distribution grids. Although these DERs provide opportunities for low-carbon energy and innovation, they can also strain the power grids. However, if properly managed, DERs can leverage their flexibility—the ability to adjust production or consumption across time and location—to support a more stable and efficient power grid. This thesis focuses on solving key challenges to unlock this flexibility.

A major challenge comes from the unpredictable nature of DERs. Renewable energy depends on the weather, and energy use of HPs and data centres varies with human behaviour and needs, making it difficult to ensure reliable use of DER flexibility. However, existing mathematical methods to deal with uncertainty are complicated and computationally slow to implement for real-world problems. This raises the first question: Q1) how can we use the DER flexibility reliably and in a computationally efficient way?

Beyond technical challenges, understanding the impacts of DERs is also important for the roadmap design to realise DER flexibility, such as impacts on power systems, economics, and society. Despite extensive research on the impacts of various types of DERs, HPs and data centres remain underexplored. In particular, the impact of HPs on fuel poverty remains insufficiently addressed. Fuel poverty is a status where households struggle to meet the fundamental heating requirement, representing a critical economic and social issue worldwide. Since HPs are around three times more efficient than traditional electric heaters, they could offer a cheaper and greener solution for homes that currently rely on expensive fuels like heating oil or older electric systems. However, it is still unclear how switching to HPs on a large scale would affect both fuel poverty levels and the costs of upgrading power distribution grids to support this change.

There is also limited understanding of the data centre impact on power systems, specifically in how much new flexibility could be unlocked to support the grid, particularly those dedicated to Artificial Intelligence (AI). With the rapid growth of the AI industry, more AI-focused data centres are being built. These AI-focused data centres are more energy-hungry than traditional data centres, but it remains unclear how much flexibility can these AI-focused data centres offer and at what cost. This ambiguity hinders the collaboration between data centres and power grids to unlock the new flexibility source. Based on all the discussions above, this

thesis answer the following additional two research questions: Q2) What are the fuel poverty and distribution grid impacts of HPs if they were to replace off-gas heating? and Q3) do the AI-focused data centres have a different potential for power system flexibility compared to traditional data centres, and what is the corresponding cost of flexibility provision?

To address the first question regarding uncertainty, this thesis first investigates standard DER operation problems and proposes a novel method to tackle DER uncertainty, which can solve the problem up to 100 times faster than traditional approaches. Then, this thesis focuses on a specific DER hierarchical operation regime, known as DER aggregation. This thesis proposes a novel and computationally efficient method to tackle uncertainty in DER aggregation. The proposed method is shown to achieve a better balance between conservativeness and reliability than existing methods, and has complexity that does not scale up with the number of DERs nor with the size of the power grid, thereby facilitating its applicability to real-world applications. To address the question on HPs, this thesis proposes a novel method that enables the analysis of fuel poverty and grid upgrade cost in England and Scotland regions in the UK. Results show that HPs could reduce fuel poverty by 20% for specific UK regions, and there are regions where HPs can lead to high grid upgrade costs. These results motivate the HP deployment and also schemes to realise their flexibility.

Finally, this thesis addresses the question on data centres. Specifically, this thesis proposes a novel analysis method and examines 14 data centres to assess how they can provide power system flexibility, finding that data centres dedicated for AI can provide greater flexibility and at a 50% lower cost than traditional data centres. This insight provides motivation for both power grid operators and AI data centre operators to collaborate in future schemes aimed at realising data centre flexibility.

Overall, this thesis contributes to the understanding of how to unlock DER flexibility in future power systems. The proposed methods for efficient uncertainty integration and the evaluation of HP and data centre impacts provide valuable insights for both researchers and practitioners. The findings highlight the potential of DERs to support reliable and efficient power system operation, while also addressing critical social issues such as fuel poverty.

Acknowledgements

What a long journey!

It feels truly incredible to stand at this point, seeing my PhD thesis finally completed. I can't help but recall the day I submitted my PhD application—and even earlier, the day I formally decided to pursue a PhD. Throughout that time, I kept asking myself: *Why do I want to do a PhD?*

When I was studying for an MSc in Artificial Intelligence (AI) at the University of Edinburgh, many AI PhD students I met amusingly advised me not to pursue a PhD—especially if my ultimate goal was to earn money or move into industry (which, frankly, still sounds quite appealing to me even now). Nevertheless, I decided to go ahead because I did find academic research fun, interesting, and fulfilling. What makes my journey even more interesting is the switch from an AI master's program to a PhD in power systems—and, behind that, the fact that my academic background was originally in power systems. More precisely, I studied Electrical Engineering and its Automation at North China Electric Power University, a university famous for its reputation in power systems in China. I decided to switch to AI over my MSc study because I thought power systems were boring, especially after hearing about the monotonous work culture at the State Grid in China (based on partial information, admittedly). However, once I started diving into AI, I realised it could be equally unfulfilling—its black-box nature, unstable training processes, endless hyperparameter tuning, results often driven by luck, and heavily intuition-based methods all felt frustrating (although I now believe that every method has its own limitations, the key is to choose the right hammer for the specific nail). This is why, during my MSc, I took two courses on mathematical optimisation (the so-called Applied Mathematics) from the School of Mathematics, which reignited my interest in applying mathematical modelling to solve more applied work. Then, I felt a strong desire to apply my skills to real-world problems and to reconnect with the foundational knowledge I gained during my undergraduate studies. All of these drove me to pursue a PhD in power systems—and now, looking back, I'm so glad I made this choice. I truly enjoy the research I'm doing.

Throughout this journey, I've been supported by many people to whom I owe my deepest gratitude. First and foremost, I want to thank my main supervisor, Professor Thomas Morstyn. He has been an outstanding mentor, offering detailed and thoughtful guidance throughout my PhD. I have learned so much from him—not only about conducting research but also about critical engineering thinking, how to approach and solve problems, how to develop good research taste, and how to communicate and deliver ideas effectively. Importantly, he never placed undue pressure on me, which is a big reason why I found my PhD experience not tough—especially compared to others I know.

I also want to thank Dr. Chaimaa Essayeh. As my unofficial “co-supervisor”, she was incredibly helpful, especially during the early stages of my PhD. Back then, I often struggled to fully grasp the ideas Thomas was trying to convey, and Chaimaa played a vital role in bridging that gap. Her encouragement helped me gain confidence in discussions and express myself more clearly. Of course, I also want to acknowledge my other official supervisors, Professor Gareth Harrison and Dr. Wei Sun. They supported me throughout the journey by providing selfless guidance and valuable help, which enabled me to successfully navigate the various stages of my PhD—including annual reviews, thesis discussions, and the organisation of my viva.

I am deeply grateful to my family—especially my mom—who has given me selfless, unconditional support throughout this journey. Even though they may not understand the details of my research, they firmly believe that I’m doing useful and great things, which has been an irreplaceable source of comfort whenever I faced challenges.

I want to thank my friends and collaborators who have supported me both emotionally and academically. These include (in random order): Haoyuan Deng, Diewu Huang, Ángel Paredes, Yuxin Xia, Hanbin Yang, and Shangmin Guo. Although they work on different fields with different expertise, our discussions have always been smooth, insightful, and fruitful. Their influence has significantly shaped my current research taste and thinking. Though we are now spread across the world, their support transcends physical distance.

Finally, a heartfelt thank you also goes to the University of Edinburgh and the University of Oxford. The time I spent in these two beautiful cities—and at these two exceptional universities—has been priceless and will remain a cherished part of my life.

Contents

Abstract	ii
Lay Summary	v
Acknowledgements	vii
Figures and Tables	xiii
Nomenclature	xvi
1 Introduction	1
1.1 Context and motivations	2
1.1.1 Complexity of power system scheduling	2
1.1.2 Inherent uncertainty of distributed energy resources	6
1.1.3 Unrecognised impact of distributed energy resources	7
1.2 Original contributions and publications	10
1.3 Thesis structure	12
2 Literature Review	14
2.1 Power system scheduling	15
2.1.1 Scheduling of distributed energy resources	16
2.1.2 Aggregation of distributed energy resources	17
2.2 Uncertainty in power system scheduling	18
2.2.1 Dealing with uncertainty in general optimisation problems	19
2.2.2 Chance-constrained power system scheduling	23
2.2.3 Uncertainty in aggregation of distributed energy resources	26
2.3 Impacts of distributed energy resources	27
2.3.1 Impact of electric vehicles	28
2.3.2 Impact of distributed renewables	30
2.3.3 Impact of distributed energy storage	31
2.3.4 Impact of heat pumps	32
2.3.5 Impact of data centres on power system flexibility	33
2.4 Gap analysis	36
2.4.1 Gaps in managing the uncertainty of distributed energy resources through chance constraints	36
2.4.2 Gaps in the impact of distributed energy resources	37

3	Strengthened and Faster Linear Approximation of Wasserstein Chance Constraints	40
3.1	Problem formulation	43
3.1.1	Exact reformulation	43
3.1.2	Convex inner-approximations	44
3.2	Strengthened and faster linear approximation	46
3.2.1	Derivation and formulation	47
3.2.2	Theoretical properties and analysis	49
3.3	Numerical experiments	51
3.3.1	Unit commitment problem	51
3.3.2	Bilevel storage strategic bidding problem	55
3.4	Concluding remarks	59
4	Aggregating Distributed Energy Resources under Uncertainty	61
4.1	Problem definition	63
4.1.1	Model of distributed energy resources	64
4.1.2	Network model	66
4.1.3	Problem target: jointly reliable AFAPR	68
4.2	Core framework of the proposed method	71
4.2.1	Surrogate polytope in a deterministic setting	71
4.2.2	Extension to jointly reliable AFAPR approximation	72
4.3	Part I: integrating WDRJCC	73
4.3.1	Exact reformulation	74
4.3.2	Cost and disaggregation	76
4.3.3	Overall framework of Part I	77
4.4	Part I: case study	80
4.4.1	Network and DER settings	80
4.4.2	Benchmark models	82
4.4.3	Setting of the proposed method	85
4.4.4	Visualisation of the AFAPR	86
4.4.5	Evaluation of the whole AFAPR	87
4.4.6	Scheduling evaluation	88
4.4.7	Sensitivity analysis	90
4.4.8	Scalability test	93
4.4.9	Computing time of the AFAPR calculation	94
4.5	Part II: integrating classical JCC	94
4.5.1	Proposed solution to joint chance constraint	94
4.5.2	Overall framework of Part II	98
4.6	Part II: case study	99

4.6.1	Case study settings	99
4.6.2	Out-of-sample reliability results	100
4.6.3	Out-of-sample cost results	101
4.6.4	Scalability	101
4.6.5	Overall computing time	102
4.7	Discussion	103
4.7.1	Comparison with the two-stage RO framework	103
4.7.2	Limitation	104
4.7.3	Potential solution	105
4.8	Concluding remarks	105
5	Heat Pump Impact on Fuel Poverty and Distribution Networks	106
5.1	Methods	107
5.1.1	Estimation of fuel poverty	108
5.1.2	Estimation of greenhouse gas emissions	111
5.1.3	Estimation of distribution network upgrade costs	112
5.1.4	Summary of assumptions	118
5.2	Results	120
5.2.1	Fuel poverty estimation	120
5.2.2	Fuel poverty reduction by heat pumps	121
5.2.3	Network upgrade cost	128
5.2.4	Cost-benefit analysis	130
5.3	Discussion	132
5.4	Limitations of study	136
5.5	Concluding remarks	137
6	Evaluating the Grid Flexibility and the Associated Cost of Data Centres	138
6.1	Methods	140
6.1.1	Evaluating the maximum amount of flexibility	140
6.1.2	Data centre cost of flexibility	146
6.1.3	Dynamic quota	151
6.1.4	Data processing	153
6.1.5	Problem settings	155
6.1.6	Summary of assumptions	155
6.2	Results	156
6.2.1	Maximum amount of data centre flexibility	156
6.2.2	Cost of data centre flexibility	159
6.2.3	Generalisability of the finding	163
6.2.4	Impact of job patterns on the maximum flexibility and flexibility cost	166
6.2.5	Dynamic quota for greater flexibility and lower cost	166

CONTENTS	xii
6.2.6 Discussion on Cooling Systems	168
6.3 Conclusion	168
7 Conclusion	170
7.1 Summary of the thesis	170
7.2 Future work	173
7.2.1 Scheduling distributed energy resources with a multi-stage uncertainty structure	173
7.2.2 Understanding the potential and value of heat pump flexibility	175
7.2.3 Realising power system flexibility of AI-focused data centres	175
Appendices	
A Appendix for the Strengthened and Faster Linear Approximation	176
A.1 Strengthened and exact MIP reformulation	176
A.2 Proofs	177
A.3 Supplementary information for the unit commitment problems	184
A.3.1 Case study settings for the unit commitment problem	184
A.3.2 Supplementary numerical experiments of the unit commitment problem	186
A.4 Supplementary information for the bilevel bidding problem	187
A.4.1 Case study settings for the bilevel bidding problem	187
A.4.2 Supplementary analysis for the bilevel bidding problem	188
B Appendix for the Fuel Poverty and Network Impact of Heat Pumps	190
B.1 Key resources table	190
B.2 Resource availability	190
B.2.1 Data and code availability	190
B.3 Amortisation and levelised cost	193
B.4 Customers' repayment towards the network cost	194
C Appendix for the Grid Flexibility and Cost of Data Centres	195
C.1 Cost minimisation problem tightening	195
C.2 Cost minimisation problem tightening under dynamic quota	196
C.3 Discussions of job partition	197
C.4 Supplementary information	197
C.4.1 Computing prices and model specifications from cloud platforms . . .	197
C.4.2 Computing speed information	202
C.4.3 Data file availability	202
C.4.4 Data file description	202
Bibliography	204

Figures and Tables

Figures

1.1	The estimated (by NESO) number of EVs and HPs in the three pathways that achieve the net-zero emission target in the UK.	4
1.2	Illustration of the aggregation process.	5
2.1	Illustration of ambiguity sets in distributionally robust optimisation. The green areas represent the ambiguity sets. Assuming all the distributions in this figure have zero mean values. (a) An example of a moment-based ambiguity set, which includes all distributions that share the same zero mean as the reference distribution. Some distributions with implausible shapes are included, which can lead to over-conservativeness. (b) An example of a distance-based ambiguity set, which includes distributions within a distance θ of the reference distribution. The two distributions with implausible shapes are excluded—even though they share the same mean as the reference distribution.	21
2.2	Future annual demand of typical DERs in GB.	28
2.3	Future annual generation of RES connected to distribution networks in GB. . . .	29
3.1	Two possible cases when a strengthening inequality valid for the original feasible region is applied to the convex inner-approximation.	42
3.2	The derivation flow and the formulation comparisons of the x -feasible region defined by different reformulations of RHS-WDRJCC.	47
3.3	Computation time to obtain the first comparable high-quality solution for the proposed SFLA and the benchmark ExactS for the UC problem.	53
3.4	Performance comparison of the optimality for the proposed SFLA and the benchmark ExactS for the UC problem.	53
3.5	Performance comparison for the bilevel problem.	57
4.1	The scheduling process using the AFAPR.	62
4.2	The PQ capability chart for (a) PV, (b) WT, and (c) ES.	66
4.3	Illustration of AFAPR and its approximation.	70
4.4	DER locations in the IEEE-123 distribution network.	81
4.5	Nodal voltages for all phases at all nodes under an unregulated baseline profile. .	82
4.6	Histogram of the real-world wind and solar forecasting error.	82
4.7	Visualisation of VB parameters of 5 of the $\tilde{\mathbf{b}}^{\text{VB}}$ samples in $\mathcal{D}$ in the ‘LowRES’ case.	86

4.8	The projected 10,000 Monte-Carlo uniform samples of the proposed MILP-based AFAPR in (4.33).	86
4.9	Evaluation for the whole AFAPR in conservativeness and inner-approximation.	87
4.10	Evaluation for minimised cost (optimality) and the out-of-sample reliability.	89
4.11	Successfully disaggregated DER power and nodal voltages (by Z-Bus nonlinear power flow) for the next $T=24$ scheduling horizon.	89
4.12	Sensitivity analysis for the out-of-sample reliability of the proposed method for different ES sizes.	90
4.13	Sensitivity analysis for the out-of-sample costs of the proposed method for different ES sizes.	91
4.14	Sensitivity analysis for the out-of-sample costs and reliability of the proposed method for different number of samples.	92
4.15	Out-of-sample costs, reliability, and scheduling time under various DER numbers.	93
4.16	Out-of-sample reliability evaluation.	100
4.17	Out-of-sample cost evaluation.	101
4.18	Out-of-sample evaluation for cost and reliability with different number of DERs.	102
4.19	Computing efficiency evaluation for the hyperplane algorithm with different number of DERs.	103
5.1	Fuel poverty estimations and proportions of off-gas households.	107
5.2	Flowchart for estimating the PS upgrade cost.	113
5.3	Breakdown of load related expenditure by UK DNOs.	118
5.4	Unit heating costs of different heating options.	123
5.5	Proportion of out-of-fuel-poverty households considering the HP upfront costs.	125
5.6	Regional fuel poverty reduction, fuel cost reduction, GHG reduction, and network upgrade cost when HPs replace off-gas heating.	127
5.7	Fuel poverty (FP) reduction and GHG emission reduction versus network upgrade cost.	129
5.8	Correlation plot for the central dimensions for the analysis of HPs on fuel poverty, network upgrade costs, and GHG emissions.	133
6.1	Baseline utilisation time series and maximum amount of flexibility for two selected data centres.	158
6.2	Average flexibility cost and power system service prices.	160
6.3	Samples of the cost scaling factor and the elements (A , R , and G) involved in its computation (see Eq. (6.21)).	161
6.4	Average flexibility cost under the cost scaling factor derived by Lambda GPU Cloud data.	162
6.5	Analysis of all the 14 HPC data centres.	164
6.6	Correlation analysis of the flexibility results and data centre utilisation patterns.	165

6.7	The maximum amount of flexibility and the flexibility cost with the dynamic quota opportunity.	167
A.1	Comparison of original samples and generated samples for wind forecasting error.	184
A.2	Performance comparison of the proposed SFLA with benchmarks for the UC problem: ExactS, LA, W-CVaR, and Bonferroni for the UC problem.	185
A.3	Computation time to obtain the first comparable high-quality solution under different optimisation horizons T for the proposed SFLA and the benchmark ExactS for the UC problem.	185
A.4	Comparison of optimality for different optimisation horizons T for the proposed SFLA and the benchmark ExactS for the UC problem.	186
C.1	The maximum amounts of flexibility calculated by solving optimisation problems with varying horizons (number of days).	198
C.2	The normalised maximum amount of flexibility (Norm. flexibility) versus the average flexibility cost when providing 100% of the maximum flexibility (ave. flex. cost) for two power system services.	199
C.3	Correlation analysis of the flexibility results and data centre utilisation patterns.	200
C.4	Baseline utilisation time series of 14 data centres before and after job aggregation.	201

Tables

1.1	List of Publications Included in the Thesis	12
3.1	Performance comparison of the final solution for the UC problem.	54
5.1	Fuel Poverty Rate (%) Estimation	121
5.2	Fuel Poverty Rate by Region	122
6.1	Summary of the analysed HPC data centres	157
A.1	Comparison of the computation time for the bilevel problem	189
B.1	Key Resources Table	191

Nomenclature

Acronyms

ACOPF	Alternating-current optimal power flow.
ACPF	Alternating-current power flow.
ADMM	Alternating direction method of multipliers.
AFAPR	Aggregated feasible active power region.
AFPR	Aggregated feasible power region.
AI	Artificial Intelligence.
BEIS	Department for Business, Energy & Industrial Strategy, which existed until 2023 when it was split to form the Department for Business and Trade (DBT), the Department for Energy Security and Net Zero (DESNZ) and the Department for Science, Innovation and Technology (DSIT). Responsibility for national security and investment policy has gone to the Cabinet Office.
CC	Chance constraint.
CCED	Chance-constrained economic dispatch.
CCP	Chance-constrained programming.
CCUC	Chance-constrained unit commitment.
CDF	Cumulative distribution function.
CL	Controllable loads.
CPU	Central processing unit.
CVaR	Conditional value-at-risk.
DCOPF	Direct-current optimal power flow.
DCPF	Direct-current power flow.
DER	Distributed energy resource.
DNO	Distribution network operator.
DRCC	Distributionally robust chance constraint.
DRCCP	Distributionally robust chance-constrained programming.
DRICC	Distributionally robust individual chance constraint.
DRJCC	Distributionally robust joint chance constraint.
DSO	Distribution system operator.
ED	Economic dispatch.
EDH	Electric direct resistive heating.
EHS	English Housing Survey.
ERCOT	Electric Reliability Council of Texas.
ERH	Electric resistive heating.

ES	Energy storage.
ESH	Electric storage heating.
EV	Electric vehicle.
GHG	Greenhouse gas.
GPU	Graphics processing unit.
HP	Heat pump.
HPC	High-performance computing.
ICC	Individual chance constraint.
IZ	Intermediate zone.
JCC	Joint chance constraint.
KKT conditions	Karush-Kuhn-Tucker conditions.
LA	Linear approximation.
LHS	Left-hand-side.
LLM	Large language model.
LoA	Local authority.
LPG	Liquefied petroleum gas.
LRE	Load related expenditure.
MILP	Mixed-integer linear programming.
MIP	Mixed-integer programming.
MISO	Midcontinent Independent System Operator.
MSOA	Middle layer super output area.
NL	Non-controllable loads.
NYISO	The New York Independent System Operator.
OPF	Optimal power flow.
PCC	Point of common coupling.
PDF	Probability density function.
PF	Power factor.
PS	Primary substation.
PV	Photovoltaic unit.
QoS	Quality of service.
RES	Renewable energy sources.
RHI	Renewable Heat Incentive.
RHS	Right-hand-side.
RHS-CC	Chance constraint with right-hand-side uncertainty.
RHS-ICC	Individual chance constraint with right-hand-side uncertainty.
RHS-JCC	Joint chance constraint with right-hand-side uncertainty.
RHS-WDRJCC	Wasserstein distributionally robust joint chance constraint with right-hand-side uncertainty.
RO	Robust optimisation.

RS	Random sampling.
SAA	Sample average approximation.
SFLA	Strengthened and faster linear approximation.
SHCS	Scottish House Condition Survey.
SLA	Service level agreement.
TSO	Transmission system operator.
UC	Unit commitment.
VB	Virtual battery
W-CVaR	Worst-case conditional value-at-risk.
WDRCC	Wasserstein distributionally robust chance constraint.
WDRICC	Wasserstein distributionally robust individual chance constraint.
WDRJCC	Wasserstein distributionally robust joint chance constraint.
WT	Wind turbine.
Functions	
$[Z]$ with $Z \in \mathbb{Z}^+$	Create a index set $\{1, \dots, Z\}$ for a positive integer Z .
$\lceil \cdot \rceil$	The round function that returns the nearest integer.
$ \cdot $	Returns the number of elements of a set, or returns the element-wise absolute values for real numbers, or returns magnitudes for complex numbers.
$\overline{(\cdot)}$	Return the element-wise complex conjugate.
$CLIP(\cdot)$	A function that sets negative elements to zero.
$C'(\cdot)$	The aggregated DER cost function $C(\cdot)$, which maps the aggregated power to the corresponding minimum total DER operational cost.
$C(\cdot)$	The total DER operational cost function, which depends on all DER decision variables (e.g., charging/discharging, load adjustments).
$\text{diag}(\cdot)$	Constructs a square matrix with the elements of the vector inside as its diagonal.
$\text{dist}(\xi_i, \mathcal{S}(\mathbf{x}))$	The distance of the i -th realisation of the random vector ξ_i from the complement of the safety set $\mathcal{S}(\mathbf{x})$.
$\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))$	The proposed conservative approximation of the distance function $\text{dist}(\xi_i, \mathcal{S}(\mathbf{x}))$.
$d_W(\mathbb{P}, \mathbb{P}')$	The Wasserstein distance between two probability distributions $\mathbb{P}$ and $\mathbb{P}'$.
$F_\xi(\cdot)$	The CDF of a random vector ξ .
$F_\xi^{-1}(\cdot)$	The inverse CDF of a single random variable ξ .
$\inf(\cdot)$	The infimum of a set, which is the greatest lower bound of that set. When the minimum exists, the infimum is equal to the minimum.
$\mathcal{J}(\cdot)$	The cost function in the upper-level system scheduling associated with decision variables $\mathbf{p}^{\text{SO}}$.

$\mathbb{P}(\cdot)$	The probability function returning the probability of the event inside $(\cdot)$, associated with a specific distribution for the random variable (vector) inside $(\cdot)$.
$\mathbb{P}\text{-CVaR}_{1-\epsilon}(\cdot)$	The CVaR of $(\cdot)$ under a specific probability function $\mathbb{P}$ at a risk level ϵ .
$\mathbb{P}_N(\cdot)$	The probability function associated with the empirical distribution for the random variable (vector) inside $(\cdot)$. The empirical distribution is constructed by assigning uniform probability mass to each historical data in the dataset.
$\sup(\cdot)$	The supremum of a set, which is the least upper bound of that set. When the maximum exists, the supremum is equal to the maximum.
∇	The vector differential operator.
$\tau^P(\cdot)$	A penalty function (e.g., penalizing voltages near operational limits) used in the disaggregation phase or feasibility check.

Indices

i	The index set for the samples of random variable/vector.
i_{CL}	The CL index.
i_{ES}	The ES index.
i_{PV}	The PV index.
i_{WT}	The WT index.
i_d	The index for the type- d DER, where $d \in \{\text{PV}, \text{WT}, \text{ES}, \text{CL}\}$.
p	The index of the constraint in JCC or WDRJCC.
t	The time step index.
ψ	The phase connection index.

Parameters

A^{PA}	The coefficient matrix that controls the slope of each facet of the polytope approximation to AFPR.
A^{VB}	The coefficient matrix that controls the slope of each facet of the VB approximation to AFPR. A VB approximation is a special case of polytope approximation.
A_j/A	The price reduction coefficient.
$\mathbf{b}^{\text{PA}}$	The coefficient vector that controls the intercept of each facet of the polytope approximation to AFPR.
$\mathbf{b}_{\text{error},i}^{\text{PA}}$	The i -th sample of the random vector $\tilde{\mathbf{b}}_{\text{error}}^{\text{PA}}$.
$\mathbf{b}_{\text{fore, hist},i}^{\text{PA}}$	The value of $\mathbf{b}^{\text{PA}}$ based on DER actual values over the i -th historical day.
$\mathbf{b}_{\text{fore, hist},i}^{\text{PA}}$	The value of $\mathbf{b}^{\text{PA}}$ based on DER forecasts over the i -th historical day.

$\mathbf{b}^{\text{VB}}$	The coefficient vector that controls the intercept of each facet of the VB approximation to AFPR. A VB approximation is a special case of polytope approximation.
$\mathbf{b}_{\text{fore}}^{\text{PA}}$	The deterministic part (forecast) of the random vector $\tilde{\mathbf{b}}^{\text{PA}}$.
$\hat{C}_{\text{elec}}$	The unit (per kWh) electricity cost.
$\hat{C}_{\text{heat}}$	The unit (per kWh) heating cost of a given primary heating option.
$\hat{C}_{\text{PS}}$	The unit network upgrade cost per exceeding MVA (£).
C_{UH}	The upfront cost of a HP (£).
$D_{P,i}$	The one-year existing electric active power demand profile for the i -th PS.
$D_{Q,i}$	The one-year existing electric reactive power demand profile for the i -th PS.
$\hat{D}_{\text{HP}}$	The normalised annual electricity demand profile of a single Heat Pump (kWh).
$\bar{D}_i$	The existing peak demand (MVA) of the i -th PS.
D_j	The computing workload for job completion, which is the smallest number of time steps required to complete the job (computing time) under the user-specified number N_j^{R} of computing resources.
$\bar{e}_{i_{\text{ES}}}$	The maximum energy storage of the i_{ES} -th ES.
$\underline{e}_t^{\text{VB}}, \bar{e}_t^{\text{VB}}$	The minimum and the maximum energy levels of VB at time step t , which is part of the entries of $\mathbf{b}$.
$\underline{e}_{i_{\text{ES}}}$	The minimum energy storage of the i_{ES} -th ES.
E_{heat}	The mean household heating energy (kWh).
G	The increase of power (kW) of an AI-focused (respectively, general-purpose) data centre when the utilisation of a single GPU (respectively, CPU) increases from 0% to 100%, which can include the contribution from memory, cooling, etc.
G_0	The parameter that captures the fixed power (kW) independent of computing, such as the lighting power or the server idle power, of the whole data centre.
H	The coefficient matrix for the three-phase AC power flow equations.
I	The identity matrix.
$J_t, j_t^0, K_t, k_t^0, g_t, g_t^0$	Coefficients for linearised three-phase power flow model.
k	A number defined as $k := \lfloor \epsilon N \rfloor$, which is introduced for deriving the strengthening inequality.
K^{DQ}	The speed-up coefficient in the dynamic quota opportunity.
M	A large constant used in mixed-integer programming formulations.
M^{P}	The time overhead triggered by each job preemption.
N	The number of collected samples of the random variable/vector ξ .

N^{CL}	The total number of controllable loads (CL).
N^{ES}	The total number of ES.
N^{NL}	The total number of non-controllable loads (NL).
N^{PV}	The total number of PV.
N^{WT}	The total number of WT.
N^B	The number of three-phase PQ buses.
N_{house}	The total number of households in a given region.
N_{LHP}	The expected HP life-span.
$\hat{N}^{\text{R}}$	The total number of GPUs (respectively, CPUs) in an AI-focused (respectively, general-purpose) data centre.
N_j^{R}	The number of GPUs (respectively, CPUs) used by job j specified by users in an AI-focused (respectively, general-focus) data centre.
$\overline{p}_{i_{\text{ES}}}^{\text{ES}, \psi}$	The <i>maximum</i> (positive) charging-power limit for the i_{ES} -th ES at phase ψ .
$\overline{p}_{t, i_{\text{ES}}}^{\text{ES}, \psi}$	The maximum active power at time step t and phase ψ for the i_{ES} -th ES.
$\underline{p}_t^{\text{VB}}, \overline{p}_t^{\text{VB}}$	The minimum and the maximum power of VB at time step t , which is part of the entries of $\mathbf{b}$.
$\underline{p}_{i_{\text{ES}}}^{\text{ES}, \psi}$	The <i>minimum</i> (negative) discharging-power limit for the i_{ES} -th ES at phase ψ .
P	The number of constraints in a WDRJCC.
p_t^{base}	The baseline data centre electric power.
q_p	The $(k+1)$ -th smallest value among the set $\{\mathbf{b}_p^\top \boldsymbol{\xi}_i\}_{i \in [N]}$.
R	The hourly computing price of a single GPU/CPU.
r_{dis}	The discount rate for amortisation or levelised cost analysis.
S	The amount (kW) of flexibility a data centre aims to provide.
$S_{\text{F}, i}$	The firm capacity of the i -th PS in MVA.
T	The total number of time steps of the optimisation problem.
t_j^{C}	The completion time of job j under the baseline scheduling strategy before providing flexibility.
t_j^{S}	The submission time of job j .
$\mathbf{w} \in \Delta_{++}$	A tunable hyperparameter that affects the performance of the W-CVaR approximation by prioritising specific constraints in a RHS-WDRJCC. Here this thesis has defined $\Delta_{++} := \{\mathbf{w} \in (0, 1)^P : \sum_{p \in [P]} w_p = 1\}$.
$\mathbf{Y}$	The complex three-phase bus admittance matrix.
$\mathbf{Y}_{00}$	The self-admittance matrix of the slack bus.
$\mathbf{Y}_{0L}$	The mutual-admittance matrix between the slack bus and non-slack buses.

Y_{L0}	The mutual-admittance matrix between non-slack buses and slack bus.
Y_{LL}	The self-admittance matrix of non-slack buses.
α^{CL}	The power factor of CL.
α^{NL}	The power factor of NL.
$\alpha_{\text{PS},i}$	The PF of the i -th PS.
κ	The hyperparameter of the proposed SFLA.
Λ	A lower-triangle matrix with all non-zero elements equal to 1.
ξ_i	The i -th sample of the random vector ξ .
Δt	The length of a single time step.
ϵ	The risk level for the CC.
ϵ_p	The risk level of the p -th ICC in the Bonferroni approximation to a JCC with P constraints.
μ_{HP}	The Seasonal Performance Factor (SPF) of a Heat Pump.
ω^{CL}	The controllable-load per-kWh cost coefficient (e.g. load-curtailment penalty).
ω^{ES}	The ES per-kWh cost coefficient (e.g. battery-degradation or charging/discharging cost).
$\omega_{i\text{ES}}$	A penalty coefficient on the i_{ES} -th ES's throughput to discourage simultaneous charge and discharge.
σ_{max}	The maximum proportion of baseload that the CL can be adjust to.
σ_{min}	The minimum proportion of baseload that the CL can be adjust to.
θ	The Wasserstein radius defining the ambiguity set.
$\tilde{\eta}$	The ES discharging efficiency.
$\hat{\eta}$	The ES charging efficiency.
δ^{max}	The maximum allowable time delay (%) of the job completion, relative to the computing time before providing flexibility.
$ i_{\text{min}}^L , i_{\text{max}}^L $	The minimum and the maximum allowable magnitudes of three-phase line currents.
$ v_{\text{min}} , v_{\text{max}} $	The minimum and the maximum allowable magnitudes of three-phase nodal currents.
Sets	
$[N]_p$	The index set defined as $[N]_p := \{i \in [N] \mid \mathbf{b}_p^\top \xi_i < q_p\}$, which is introduced for deriving the strengthening inequality.
$\mathbb{B}$	The domain of binaries.
$\mathbb{C}$	The domain of complex numbers.
$\mathbb{R}$	The domain of real numbers.
$\mathbb{Z}$	The domain of integers.

$\mathcal{D}$	A set of samples representing the possible realisations of the vector $\mathbf{b}$ for the polytope approximation of AFPR.
$\mathcal{D}_{\text{error}}$	A set of samples representing the possible realisations of the forecasting error of the vector $\mathbf{b}$ for the polytope approximation of AFPR.
$\mathcal{F}(\theta)$	The Wasserstein ambiguity set, which is defined as a ball with radius θ centered at the empirical distribution $\mathbb{P}_N$ constructed by historical data.
$\mathcal{J}$	The index set of computing jobs.
$\mathcal{K}^0$	The initialised constraint set in the supporting hyperplane algorithm to replace the original classical JCC.
$\mathcal{K}^j$	The constraint set in the j -th iteration of the supporting hyperplane algorithm to replace the original classical JCC.
$\mathcal{K}^W$	The constraint set in the W -th (the last) iteration of the supporting hyperplane algorithm to replace the original classical JCC.
$\mathcal{N}^A$	The set that contains indices i of the activation of power system service.
$\mathcal{PQ}_{t,i_{\text{ES}}}^{\text{ES},\psi}$	The feasible active-reactive operating region for the i_{ES} -th ES at time step t and phase ψ .
$\mathcal{PQ}_{t,i_{\text{PV}}}^{\text{PV},\psi}$	The feasible active-reactive operating region (PQ chart) for the i_{PV} -th PV at time step t and phase ψ .
$\mathcal{PQ}_{t,i_{\text{WT}}}^{\text{WT},\psi}$	The feasible active-reactive operating region for the i_{WT} -th WT at time step t and phase ψ .
$\mathcal{S}(\mathbf{x})$	The safety set of CCs, which is the feasible region for the random variable/vector for specific decision variables $\mathbf{x}$.
$\mathcal{T}$	The index set of time steps $\{1, \dots, T\}$ in the optimisation horizon.
$\mathcal{T}_i^A$	The time steps to sustain flexibility over the i^{th} activation of power system service.
$\mathcal{T}_j$	The index set collecting the available time steps of job j .
$\mathcal{X}$	The compact domain for decision variables $\mathbf{x}$.
$\mathcal{X}^{\text{Agg}}$	A compact domain for the aggregated power $\mathbf{p}^{\text{Agg}}$, which can be defined by the maximum and minimum aggregated power of the network based on the installed capacity.
$\mathcal{X}^{\text{SO}}$	The feasible set for the upper-level system operator's decision variables $\mathbf{p}^{\text{SO}}$.
$\mathcal{X}_{\text{Exact}}$	The feasible space of the decision variables $\mathbf{x}$ (the $\mathbf{x}$ -feasible space) of the exact reformulation of RHS-WDRJCC.
$\mathcal{X}_{\text{LA}}$	The $\mathbf{x}$ -feasible space of the LA approximation for RHS-WDRJCC.
$\mathcal{X}_{\text{SFLA}}$	The $\mathbf{x}$ -feasible space of the proposed SFLA for RHS-WDRJCC.
$\mathcal{X}_{\text{WCVaR}}$	The $\mathbf{x}$ -feasible space of the W-CVaR approximation for RHS-WDRJCC.

Ω^{Agg}	The Aggregated Feasible Active Power Region (AFAPR), i.e. the set of all feasible $\mathbf{p}^{\text{Agg}}$ when considering DER and network constraints.
$\Omega^{\text{Agg}}(\epsilon)$	The <i>jointly reliable</i> AFAPR under risk tolerance ϵ , satisfying all constraints with probability at least $1 - \epsilon$.
$\Omega_{\text{Approx}}^{\text{Agg}}$	An inner polytope approximation of the AFAPR Ω^{Agg} in the deterministic case.
$\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$	A distributionally robust approximation to $\Omega^{\text{Agg}}(\epsilon)$ under Wasserstein radius θ .
Ψ	The index set of phases, which comprises a, b, c for Wye connections or ab, bc, ca for Delta connections.

Uncertain parameters or random variables

$\tilde{\mathbf{b}}^{\text{PA}}$	The random vector representing the intercepts of the polytope approximation to AFPR under uncertainty.
$\tilde{\mathbf{b}}_{\text{error}}^{\text{PA}}$	The random vector representing the uncertain forecasting error of the intercepts of the polytope approximation to AFPR under uncertainty.
$\ddot{p}_{t, i_{\text{CL}}}^{\text{CL}, \psi}$	The uncertain base load of the CL without any intervention.
$\ddot{p}_{t, i_{\text{NL}}}^{\text{NL}, \psi}$	The uncertain base load of the NL without any intervention.
$\overline{p}_{t, i_{\text{PV}}}^{\text{PV}, \psi}$	The uncertain maximum active power at time step t and phase ψ for the i_{PV} -th PV.
$\overline{p}_{t, i_{\text{WT}}}^{\text{WT}, \psi}$	The uncertain maximum active power at time step t and phase ψ for the i_{WT} -th WT.
ξ	The random vector used to model uncertainty.

Decision variables

$ACoF$	The average cost of flexibility.
$AECoF$	The average extra energy cost of flexibility.
$APCoF$	The $ACoF$ without dynamic quota, which is used only for the cost calculation in the dynamic quota scenario because it involves extra energy cost than the pure $ACoF$ without dynamic quota.
$\underline{C}$	The lower bound $\underline{C}$ of the total flexibility cost C^{Flex} .
C^{E}	The extra data centre energy cost in the dynamic quota scenario.
C'_{house}	The additional annual energy bill for one household.
C_{net}	The total network upgrade cost for a given region.
C_{SW}	The modelled cost for space heating and water heating in the dataset.
C_{YU}	The annual energy bill considering the HP upfront cost.
$C_{\text{PS}, i}$	The upgrade cost to accommodate the new demand of HPs for the i -th PS.
C^{Flex}	The total flexibility cost.
$\mathbf{D}_{\text{new}}$	The estimated new annual demand profile after HP replacing off-gas heating (apparent power in MVA).

$D_{EH,i}$	The demand profile of the existing electric heating demand for the i -th PS.
$D_{HP,i}$	The demand profile of all the new HPs connected to the i -th PS.
$\bar{D}_{new,i}$	The estimated new peak demand of the i -th PS after HP replacing off-gas heating (MVA).
E^f	The total shifted energy E^f in kWh over the service activation periods.
E_A	The annual electricity consumption of a HP in kWh.
E_{other}	Energy consumption other than heating, including cooking and lighting.
E_{SW}	The energy consumption in kWh for space and water heating.
$e_{t,i_{ES}}$	The stored energy at time step t for the i_{ES} -th ES.
e_j	The scheduled completion time after flexibility provision.
f_t	The flexibility at each time step t in reducing demand.
i_t^D	The phase-to-phase currents of all the PQ buses.
i_t^L	The three-phase line currents for all lines.
i_t	The phase net current injections of all the PQ buses.
LC	The levelised lifetime cost (£/kWh).
m	The decision variables for calculating the parameters of the Bonferroni approximation to RHS-WDRJCC.
N_{PS}	The estimated total number of Primary Substations in a region.
$N_{EH,i}$	The number of existing electric heaters connected to the i -th PS.
$N_{house,i}$	The number of households connected to the i -th PS.
$N_{HP,i}$	The number of required HPs to replace all the off-gas heaters, or equally the number of off-gas heaters connected to the i -th PS.
n_j^P	The total amount of job preemption of job j .
p^{SO}	The upper-level system operator's decision variables other than the aggregated power of an aggregator p^{Agg} .
$\tilde{p}_{t,i_{ES},\psi}^{ES,\psi}$	The discharging power at time step t and phase ψ for the i_{ES} -th ES.
$\hat{p}_{t,i_{ES},\psi}^{ES,\psi}$	The charging power at time step t and phase ψ for the i_{ES} -th ES.
p_t^{Agg}	The aggregated active-power injection at time step t (summed over all phases).
$p_{t,i_{ES}}^{ES,\psi}$	The scheduled active power at time step t and phase ψ for the i_{ES} -th ES.
$p_{t,i_{PV}}^{PV,\psi}$	The scheduled active power at time step t and phase ψ for the i_{PV} -th PV.
$p_{t,i_{WT}}^{WT,\psi}$	The scheduled active power at time step t and phase ψ for the i_{WT} -th WT.
$p_{t,i_{CL}}^{CL,\psi}$	The scheduled active power at time step t and phase ψ for the i_{CL} -th CL.

p_t	The scheduled data centre electric power.
$q_{t,i_{ES}}^{ES,\psi}$	The scheduled reactive power at time step t and phase ψ for the i_{ES} -th ES.
$q_{t,i_{PV}}^{PV,\psi}$	The scheduled reactive power at time step t and phase ψ for the i_{PV} -th PV.
$q_{t,i_{WT}}^{WT,\psi}$	The scheduled reactive power at time step t and phase ψ for the i_{WT} -th WT.
$q_{t,i_{CL}}^{CL,\psi}$	The scheduled reactive power at time step t and phase ψ for the i_{CL} -th CL.
$\mathbf{r}$	The ancillary variables for the reformulation of RHS-WDRJCC.
s_t^{Agg}	The aggregated three-phase complex power.
s_t^D	The three-phase complex power injection of all the PQ buses from Delta sources.
s_t^Y	The three-phase complex power injection of all the PQ buses from Wye sources.
s	The ancillary variable for the reformulation of RHS-WDRJCC.
$S_{E,i}$	The exceeding capacity of the i -th PS in MVA.
s_i	The amount of the provided flexibility over the i^{th} activation of flexibility.
t^{Ex}	The time delay of job j after providing flexibility.
$\mathbf{v}_t^0$	The three-phase complex nodal voltage of the slack bus.
$\mathbf{v}_t$	The three-phase complex nodal voltages of all the PQ buses.
W^{delay}	The total delayed computing workload of all jobs.
$\mathbf{x}$	The vector of decision variables in a general CCP/DRCCP.
$\mathbf{x}_t^{\text{DER}}$	A vector collecting individual DER decision variables (and possibly uncertain parameters such as NL) at time step t .
$x_{j,t}$	The completed proportion of workload of job j at time step t .
$x'_{j,t}$	The binary variable that models the job running status at time step t .
$\mathbf{z}$	The ancillary binary variables used in MIP formulations of RHS-WDRJCC.
$z_{j,t}$	The preemption counter variable that quantifies the amount of preemption for job j at time step t .
α^B	The decision variables for calculating the parameters of the Bonferroni approximation to RHS-WDRJCC.
α^W	The ancillary decision variables for the W-CVaR reformulation of RHS-WDRJCC.
β^B	The decision variable for calculating the parameters of the Bonferroni approximation to RHS-WDRJCC.
β^W	The ancillary decision variable for the W-CVaR reformulation of RHS-WDRJCC.

η	The decision variable for calculating the parameters of the Bonferroni approximation to RHS-WDRJCC.
τ	The ancillary decision variable for the W-CVaR reformulation of RHS-WDRJCC.
δ_j	The delay proportion for job j .

Introduction

The modern power system is undergoing a transformative shift, driven by the rapid integration of distributed energy resources (DERs). This thesis defines DERs as energy resources in/close to the demand side or the grid edge, which can thus include distributed renewable energy sources (RES) such as solar photovoltaic units (PVs) and wind turbines (WTs), heat pumps (HPs), electric vehicles (EVs), distributed energy storages (ES), and even data centres (the computing demand side). The philosophy behind our definition aligns with that used by Ofgem (the UK energy regulator), which defines distributed generation as “generation close to where the energy is used” [1]; however, we do not additionally constrain DERs to be “small-scale”, as is sometimes done in the literature [2]. The DER integration is surging due to advancements in technology, policy incentives, and the global push towards decarbonisation [3–6].

DERs can place significant strain on power grid capacity, especially at the distribution grid level, by increasing demand or causing reverse power flows, thereby necessitating expensive network upgrades to maintain reliable operation. Fortunately, these resources often have “flexibility”, which is the ability to flexibly adjust the power demand/generation in response to the power grid signals. Properly using the DER flexibility can minimise network upgrade costs while providing valuable ancillary services to support stable grid operation. The International Energy Agency’s 3DEN initiative estimates smart flexibility could reduce global network reinforcement costs by USD \$270bn to 2040 [7]. Therefore, harnessing the DER flexibility is crucial for the development of cost-effective and sustainable energy systems [8, 9].

Unlocking DER flexibility requires several critical challenges to be addressed. These challenges span technical aspects such as computationally efficient DER scheduling under uncertainty, as well as non-technical aspects such as understanding the impacts of DERs on power systems, the economy, and society, which are equally important to support the design and implementation of effective roadmaps to realise DER flexibility. Resolving these challenges lies at the heart of this thesis.

This introductory chapter outlines the background, motivation, and key contributions of this thesis. In what follows, Section 1.1 provides a detailed discussion of the background and research questions, while Section 1.2 summarises the primary contributions and the associated publications. Finally, Section 1.3 presents an overview of the thesis structure.

1.1 Context and motivations

DERs possess unique characteristics compared to traditional large-scale, centralised resources. These include their large number, distributed placement across the network, and inherent operational uncertainty due to factors such as weather conditions and human behaviour (this includes the uncertain computing demand in data centres). Moreover, there remains a general lack of familiarity among end-users and stakeholders regarding the value and impacts of DERs. These characteristics not only exacerbate existing challenges in power system operation but also introduce new challenges as power systems move toward fully harnessing the value of DERs.

1.1.1 Complexity of power system scheduling

Many DERs have the potential to flexibly adjust their power generation and/or demand. For example, the charging of EVs and heating of HPs can be delayed to less congested periods. The computing jobs in a data centre can also be delayed or even transferred to another data centre in a less congested grid area. Even non-dispatchable RES can flexibly adjust the reactive power due to the functionality of inverters [10]. This flexibility can be leveraged to provide power system services that are critical for maintaining stable and reliable system operation [11]. To fully exploit DER flexibility, theoretically, one can solve a mathematical optimisation problem that incorporates all DERs, such that the objective is optimised while operational constraints are respected (with sufficient communication devices and control functionality in place). These flexibility optimisation problems are essentially similar to traditional power system scheduling problems, which aim to determine the most cost-efficient power generation levels for each generator across each time step while satisfying power network constraints [12]. In the context of flexibility optimisation, additional constraints that model DER flexibility can be incorporated on top of the traditional scheduling formulation.

Given the relevance of DER flexibility optimisation and the traditional power system scheduling problems, this thesis starts with two typical power system scheduling problems, which are the Economic Dispatch (ED) problem and the Unit Commitment (UC) problem [12].

ED is a core decision-making problem in power system scheduling [13]. Its primary objective is to determine the power output of available generators to meet electricity demand at the lowest possible operation cost. Network constraints are usually incorporated, which are typically represented using a direct-current power flow (DCPF) model for computational tractability.

A classic example of ED is the merit-order curve without the network constraints. In this approach, generators are ranked in an ascending order based on their marginal costs, and are dispatched from the lowest-cost generators until the demand is met. ED is also used in some electricity market clearing. In the same merit-order-curve example, the generator that precisely meets the electricity demand is designated as the marginal generator, and its price is the marginal price [12]. This clearing process is used in the current day-ahead market in Great Britain [14].

UC is a problem that determines the on/off statuses of generators to minimise commitment and operation costs of generators. Similar to ED, network constraints can also be incorporated. UC and ED often share a large portion of variables and constraints, such as power generation levels and network state variables. However, UC uniquely accounts for generator-specific constraints such as minimum on and off time limits and the start-up and shutdown characteristics of generation units [15, 16]. While generation levels are modelled in UC, these values are subject to adjustment. Consequently, ED is typically implemented after UC to refine generation schedules based on the fixed commitment statuses determined by UC [17].

These two problems have long posed significant challenges for power system operators. First, the UC problem is inherently complex due to the need to determine on/off statuses for generators, making it a mixed-integer programming (MIP) problem that is known to be NP-hard. Second, accurately incorporating network constraints into both the UC and ED problems requires using the alternating current power flow (ACPF) model, which is non-linear, non-convex, and thus computationally intractable [18]. In practice, the linearised DCPF model is commonly used to improve tractability [12] (Note that DCPF is usually used in transmission networks due to the higher accuracy. For distribution networks there are also linear network models). However, it was pointed out in [19] that DCPF-based solutions is never AC-feasible, and the process of recovering the constraint satisfaction can lead to sub-optimal operation costs [20]. This issue becomes particularly relevant in the era of DER integration, as these resources are predominantly connected at the distribution network level. In such networks, voltage drops, higher resistance-to-reactance ratios, and load imbalances further deteriorate the accuracy of the DCPF model.

Furthermore, even with DCPF, computational challenges persist due to the tremendous scale of modern power systems. For instance, the Midcontinent Independent System Operator (MISO), one of the largest electricity markets globally, manages a network model comprising over 45,000 buses, 1,400 generation sources, and approximately 2,500 pricing nodes, which lead to computationally demanding ED and UC [21, 22]. In addition, practical operations often include contingency constraints, such as the N-1 requirement, which ensures system reliability in the event of a single fault [23]. Such reliability requirements add another layer of complexity.

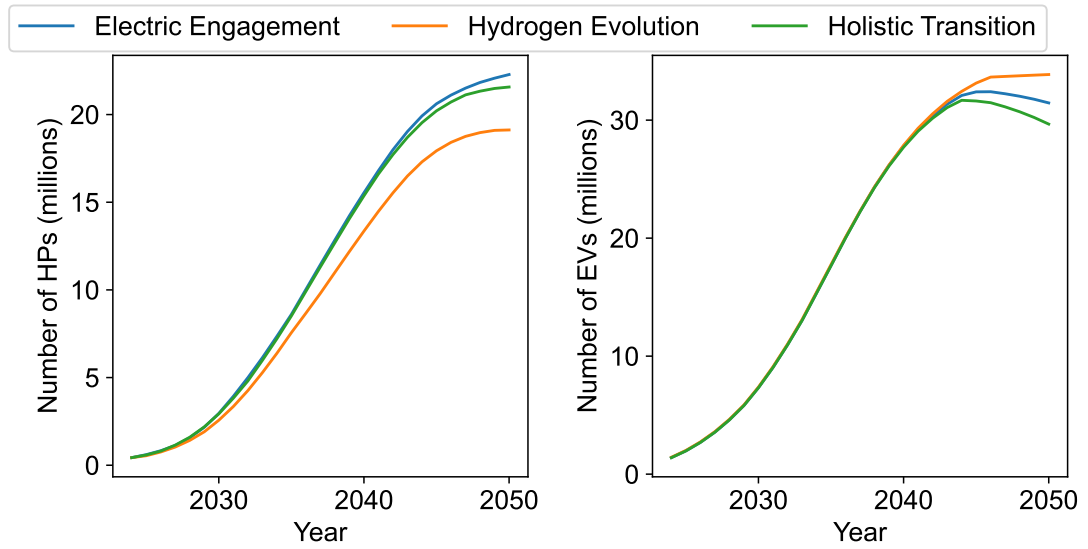


Figure 1.1: The estimated number of EVs and HPs in the three pathways that achieve the net-zero emission target in the UK by 2050 [9].

The three pathways are “Electric Engagement”, “Hydrogen Evolution”, and “Holistic Transition” as defined in [9]. For the number of EVs, the fall in numbers post 2040 is due to assumptions on other forms of transport.

In a MISO production example for the day-ahead UC problem, the MIP optimality gap is still around 40% after a computing time of 1,200 s with the CPLEX12.5 solver [21]. The surging integration of DERs further increase the problem scale and leads to greater computational intractability, as will be discussed in the following sections.

Scheduling problems of distributed energy resources

The penetration of DERs has been increasing rapidly in recent years and is expected to accelerate further [9]. DER flexibility can be leveraged for various applications, including network constraint management, reserve provision, and frequency regulation [24]; therefore, it is important to fully harness the benefits of DERs. As previously mentioned, fully utilising DER flexibility necessitates the formulation of a large-scale, whole-network optimisation problem that accounts for all DERs. However, as the number of DERs continues to grow, this large-scale problem becomes computationally intractable [25]. As illustrated in Fig. 1.1, by 2030, the number of EVs and HPs in the UK is projected (estimated by NESO) to reach several million—orders of magnitude greater than the 1,400 generation sources managed by MISO that already presents significant scheduling challenges [21, 22]. By 2050, when the UK aims to achieve its net-zero target, the number of EVs and HPs is expected to rise to 20–30 million, further exacerbating the computational demands of scheduling.

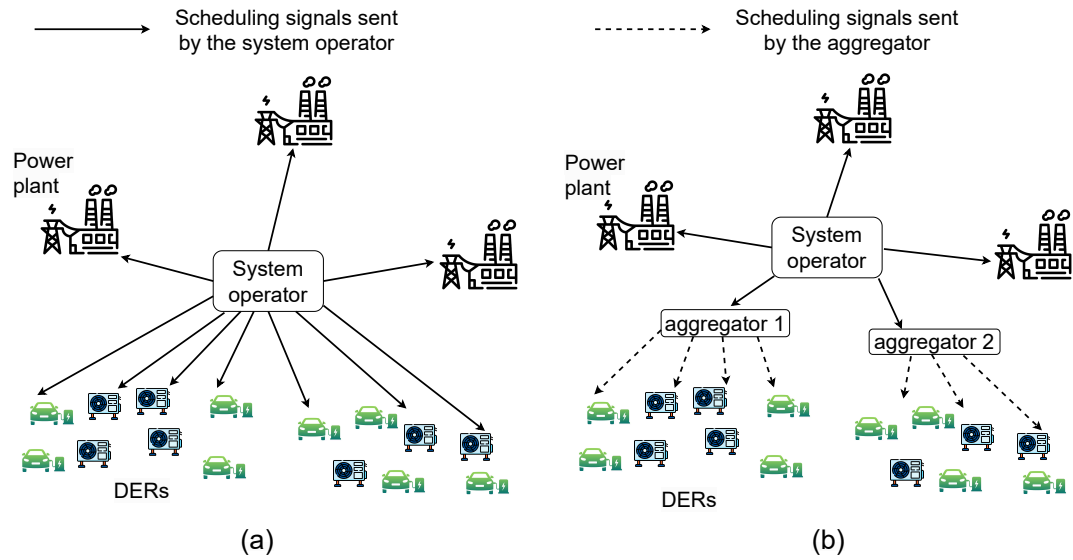


Figure 1.2: Illustration of the aggregation process.

(a) Direct system scheduling where the system operator includes all power generators and millions of DERs into its scheduling problem, which can be computationally unsolvable. (b) DERs are first aggregated by different aggregators. The system operator only needs to schedule the aggregated power of these aggregators, which leads to more manageable scheduling problem. The aggregators then schedule the DERs based on the system operator's dispatch.

One potential solution to the computational challenges of DER scheduling is distributed optimisation. In this approach, the original large-scale problem is decomposed into smaller sub-problems, which are solved independently by different entities. The optimal solution to the original problem can be found through iterative information exchange and updates between sub-problems [26]. The alternating direction method of multipliers (ADMM) algorithm is a typical example in distributed optimisation, and has found applications in DER scheduling [27,28]. However, a major issue of distributed algorithms arises from the considerable communication burden caused by the iterative exchange of information required to achieve convergence [29]. As the problem scale increases, more iterations are required to achieve convergence, increasing the communication burden [30].

Aggregation of distributed energy resources

Another promising solution is DER aggregation [29]. The core idea is to replace the original direct system-wide scheduling with a hierarchical scheduling strategy with the help of aggregators. Fig. 1.2 illustrates the hierarchical scheduling process. Numerous grid-edge devices are grouped and managed by different aggregators. Instead of directly scheduling numerous individual grid-edge devices, the system operator can schedule the aggregated power for each aggregator, while the aggregators are responsible for dispatching their own DERs—thereby alleviating computational and communication burdens.

1.1.2 Inherent uncertainty of distributed energy resources

Most DERs are inherently characterised by significant uncertainties, which arise from their dependence on external and often unpredictable factors. For instance, the output of PVs and WTs is highly influenced by environmental conditions, such as solar irradiance, cloud cover, and wind speed, all of which exhibit natural variability and are hard to predict [31]. Similarly, the operation of EVs is affected by uncertain user behaviours, particularly plug-in and plug-out times, which are hard to foresee precisely [32]. HPs also possess uncertainty, as their operation depends on uncertain outdoor temperatures and uncertain user behaviours, such as hot water usage and practices like opening windows for ventilation [33]. Finally, data centre demand fluctuates based on job arrival rates and computational requirements [34], both of which are inherently uncertain and difficult to predict.

These uncertainties pose significant challenges to the reliable operation of power systems. In traditional operations, reliability is maintained by pre-determining generation and demand quantities through long-term contracts in forward markets and refining these commitments during short-term planning, such as in day-ahead markets. The remaining imbalances are relatively minor and can be effectively managed using (near) real-time markets and power system reserves [35]. A key factor contributing to this effectiveness is the dominance of thermal generators in traditional power systems, which are dispatchable and capable of providing reliable delivery guarantees. However, current power systems have a higher penetration of renewable generation, which is less dispatchable and less reliable due to its uncertain nature. Moreover, the integration of DERs can lead to a higher level of uncertainty. In 2012, the National Renewable Energy Laboratory (NREL) [36] showed that their wind generation forecasting error could be three times that of the load forecasting error: the load forecasting error in the New York Independent System Operator (NYISO) had a normalised standard deviation of only 0.036 (the peak load was around 33,000 MW), whereas the wind forecasting error in the Electric Reliability Council of Texas (ERCOT) exhibited a normalised standard deviation of 0.119 (the corresponding wind power capacity was around 9,000 MW). On the demand side, new DERs such as EVs, HPs, and data centres introduce additional sources of variability, further increasing system-wide uncertainty. Consequently, integrating robust uncertainty management strategies into system operations becomes crucial to maintaining reliable power system operation in the face of escalating uncertainty levels.

This thesis particularly focuses on uncertainty in constraints, which poses a significant challenge for ensuring the reliable operation of power systems. Solution approaches to this challenge typically fall into one of two categories: 1) robust optimisation (RO), which guarantees constraint satisfaction under all possible scenarios, and 2) chance constraints (CC), which ensure a specified probability of constraint satisfaction [37]. In practical applications, joint chance constraints (JCC) are often preferred due to the guarantee of the simultaneous satisfaction of multiple constraints, which provides higher reliability [38]. However, RO, CC, and particularly

JCC approaches introduce significant computational complexity and can lead to non-convex and intractable formulations [37, 39]. This complexity is compounded by the fact that existing power system operational optimisation problems are already large-scale and potentially non-convex, as has been discussed in Section 1.1.1. The importance of uncertainty management, and the associated computational challenges, motivates the first research question of this thesis:

Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering the DER uncertainty?

More specifically, this thesis will address both general power system DER scheduling problems and exploring DER aggregation—given its promising performance in efficiently managing the large-scale integration of DERs, as introduced in Section 1.1.1.

1.1.3 Unrecognised impact of distributed energy resources

Beyond the technical challenge of scheduling DERs, existing work has investigated the technical, economic and social impacts of many DERs. These studies also represent a crucial dimension that supports the design of effective and efficient roadmaps to realise DER flexibility. Many existing studies have focused on EVs, distributed RES, and distributed ES, where a wide spectrum of their impacts on the power networks, economy, and society has been investigated. However, there are still two important types of DERs that remain under-investigated, which are 1) HPs, which are key to heating electrification, and 2) data centres, which are becoming more widespread and energy-demanding due to the surge of Artificial Intelligence (AI). Given the pivotal role that HPs and data centres are expected to play in future power systems, this knowledge gap poses a significant challenge in unlocking the full value of their flexibility.

Unrecognized fuel poverty impact and network impact of heat pumps

A critical but under-investigated dimension of HPs is on their impact on fuel poverty. Fuel poverty is a state where households struggle to afford the energy needed to heat their homes comfortably [40]. The far-reaching consequences of fuel poverty on households and society include physical and mental health issues [41] and increased financial demands on national health systems [42]. These impacts highlight fuel poverty as a significant and persistent social issue in the UK, where the government has been implementing various programmes to monitor and mitigate the problem over several decades [43, 44]. Recent global and political events, such as the COVID-19 pandemic and the Russia–Ukraine conflict, have exacerbated concerns surrounding fuel poverty. These events led to soaring energy prices and subsequent inflation, particularly acute during 2022 and 2023 [45, 46]. In the UK, energy bills reached

unprecedented levels during this period, with an alarming two-thirds of households at risk of experiencing fuel poverty by January 2023 [47]. Although the soaring trend of energy prices has been stopped, they have not returned to pre-crisis levels. As a result, addressing fuel poverty remains a critical challenge for social development and economic prosperity [48].

One of the most notable and promising characteristics of HPs is their roughly 300% running efficiency, which means that the same heating requirement can be met with just one-third of the energy consumption [49]. Note that, in the UK, due to gas (natural gas) being one of the major generation source for electricity (around 25% in 2024 [50]) and the marginal pricing mechanism [51], the electricity price is closely coupled with gas, and thus the HP running cost is only comparable or may be slightly more expensive than gas heaters [52, 53] (note that HPs may exploit flexible electricity tariffs to have lower cost). However, HPs can lead to lower energy cost than off-gas heating [49]. Typical off-gas heating options are electric resistive heating (ERH), heating oil, and liquefied petroleum gas (LPG), which are expensive; thus, homes with off-gas heating are associated with the highest risk in fuel poverty [43, 54], and possess urgent need for deploying lower-cost, efficient heating solutions such as HPs. Based on official data in 2023 [55], around 15.7% of UK households do not have access to the mains gas grid (i.e., are “off-grid”). In the same year, about 40% of UK homes relied on non-gas heating as their main heating method during winter [56]. This figure is higher than the 15.7% off-gas-grid share because many households used portable heaters that winter [56], likely in response to high energy bills. Therefore, there exists a non-trivial benefit of HPs in alleviating the critical fuel poverty issue especially for these off-gas homes. Understanding the impact of HPs on fuel poverty is important for promoting their residential adoption. Furthermore, a wide adoption of HPs is a fundamental step towards unlocking HP flexibility, because the value and significance of their flexibility only appear when these individually small-power-size HPs are installed beyond a certain scale, which can then provide great incentives for network operators and policy makers to push the realisation of HP flexibility.

Although the introduction of HPs may help alleviate the critical fuel poverty issue especially for off-gas homes, HPs can also add significant demand to power systems. It has been estimated that the peak grid demand in Great Britain could increase by 7.5 GW (14%) if 20% of households adopt HPs [57]. However, examining demand increase alone is insufficient to fully understand the impact of HPs on the network. This thesis specifically focusses on distribution networks, which are the networks where HPs are connected to and faces considerable challenge in the context of DER integration [58]. Electricity networks typically have a certain level of demand headroom, making it possible to accommodate the additional demand without requiring immediate action. Therefore, an important question is whether existing distribution networks can host the extra HP demand, what upgrades would be needed, and at what

cost. This evaluation would serve as an important motivation for unlocking HP flexibility, by estimating the saving of the expensive network cost if HP flexibility is activated. Taken together, the dual challenges of addressing fuel poverty and understanding the impact on the network motivate the second research question of this thesis:

Q2: What are the fuel poverty and distribution network impacts of replacing domestic off-gas heating systems with HPs?

Power system flexibility of data centres dedicated to Artificial Intelligence

Data centres are another emerging DER, but it remains under-investigated in terms of how much new flexibility can be unlocked to support the grid. Data centres house computing resources, including servers, storage systems, and network infrastructure, within controlled environments to support a wide range of applications, such as web services, data storage, and high-performance computing (HPC) services [59]. HPC is a prominent use case for data centres, where a large number of computing resources are aggregated to achieve superior computational performance capable of meeting the demands of complex applications, such as solving advanced mathematical problems, conducting physics simulations, and supporting medicinal design [60]. Traditionally, HPC applications have relied heavily on Central Processing Units (CPUs) [61].

The recent surge of AI, especially large language models (LLMs), has led to a rapid development of another type of data centres dedicated to AI development. These AI-focused HPC data centres rely on Graphics Processing Units (GPUs), and which makes them more energy-demanding than traditional general-purpose CPU-intensive HPC data centres. For example, a latest-generation “192-thread AMD EPYC 9654 CPU” has a rated power of only 360 W, while the latest-generation “NVIDIA B200 GPU” can draw 1000 W. Note that, this thesis does not compare the energy consumption per unit of computation (Watt/FLOP)—although GPUs have lower Watt/FLOP due to their strong parallelisation ability, GPUs do not support computing jobs as general as CPUs, and AI jobs tend to be more computationally intensive than traditional HPC jobs; therefore, this thesis compares the energy consumption of the single computing unit (CPU or GPU) commonly used as a reference for both traditional HPC and AI jobs.

The other contributor is the rapidly expanding use of LLMs, which requires intensive computing [62]. Venture Capital firms invested \$290 billion in AI over the last 5 years [63]. The Electric Power Research Institute (EPRI) estimates that data centres now consume 4% of the total power generation in the US, and this proportion could increase to as much as 9.1% by 2030, with AI being one of the main drivers [64]. The increased energy demand for AI poses a major challenge for power grid infrastructure [65], especially when considered alongside the ongoing efforts to electrify heating and transportation to achieve net-zero carbon emissions.

Being aware of this challenge, AI companies and hardware manufacturers are making their systems more energy efficient [66, 67]. However, Jevons Paradox suggests that increasing energy efficiency may lead to greater demand for computing and thus higher overall energy usage [68].

To mitigate the stress on power systems, data centres could be operated to provide grid flexibility. For example, computational tasks in data centres can be strategically scheduled to align with periods of high renewable generation or low system demand, thereby reducing renewable energy curtailment and supporting grid stability. Power flexibility of traditional general-purpose HPC data centres has been an active area of research for many years [69, 70]. However, as discussed, AI-focused HPC data centres are more energy-intensive due to their reliance on GPUs, which may result in different capabilities for providing power system flexibility services compared to general-purpose HPC data centres. Furthermore, recent research [71] has identified distinct job patterns between these two types of data centres, which will influence their ability to contribute to power system flexibility.

Another critical yet under-explored aspect is the quantification of the cost of data centre flexibility. Data centres are primarily built for computing, but providing power system flexibility (through adjusting computing workload) can delay the delivery of computing jobs. This creates a conflict, as data centre operators aim to complete computations as soon as possible, especially amid the current competition in AI product development. Understanding the cost of flexibility provision is important to inform data centre operators on the potential profitability of providing power grid flexibility and motivate their engagement. Cost modelling is also important for policymakers and regulators who want to assess the overall benefits of integrating data centres into power system flexibility markets. Overall, the above discussions motivate the third research question of the thesis:

Q3: Do AI-focused data centres have a different potential for grid flexibility compared to general-purpose data centres, and what is their cost of flexibility provision?

1.2 Original contributions and publications

This thesis addresses the three research questions outlined earlier and advances the understanding of uncertainty management, impact of heat electrification on fuel poverty and network upgrades, and the impact of AI-focused data centres on power system flexibility (by evaluating the maximum flexibility and the corresponding cost). The original contributions of this thesis include:

- C1 The thesis proposes a method that brings up to 90× speed-up for general DER scheduling problems with JCC. This is achieved through the proposed strengthened and faster linear approximation (SFLA) for Wasserstein distributionally robust joint chance constraints (WDRJCC) with right-hand-side uncertainty. WDRJCC addresses the limita-

tions of classical JCC by accommodating cases where the true distribution of random variables is unknown. The proposed SFLA demonstrates up to 90× speedup over traditional methods such as worst-case conditional value-at-risk (CVaR) without compromising solution quality (Chapter 3 and Appendix A).

- C2 The thesis develops a method to integrate WDRJCC and classical JCC to ensure reliable DER aggregation. The proposed method is demonstrated to achieve a superior balance between conservativeness and reliability compared to existing methods. The computational complexity of the proposed method is low, and does not increase with the number of DERs nor with the size of the network, thereby facilitating its applicability to real-world, industry-scale problems (Chapter 4).
- C3 This thesis proposes a method to analyse the fuel poverty and network impacts of HPs in the UK. The proposed method enables the integration of datasets from multiple public sources so as to carry out the combined fuel poverty and network upgrade analysis. The results show that replacing off-gas heating with HPs could lift up to 20% of households out of fuel poverty in specific UK regions, providing both social and economic incentives for HP deployment. However, this replacement would incur significant distribution network upgrade costs in certain areas, underscoring the need to realise HP flexibility (Chapter 5 and Appendix B).
- C4 The thesis proposes a novel method to evaluate both the maximum flexibility and the corresponding cost for data centres providing power system flexibility services, ranging from short-duration, low-frequency services to long-duration, high-frequency services. To ensure robust conclusions, the proposed method is applied to datasets from 14 real-world data centres and three mainstream cloud computing platforms. The results demonstrate that AI-focused, GPU-intensive data centres can deliver greater flexibility at up to 50% lower cost compared to traditional CPU-intensive data centres. Furthermore, the proposed method enables to compare the data centre flexibility cost with power system flexibility prices. By comparing flexibility cost and real-world power system service prices, this thesis illustrates the financial profitability of flexibility provision for AI-focused HPC data centres (Chapter 6 and Appendix C).

The contributions of this thesis, as outlined above, collectively address the challenges associated with unlocking the flexibility of DERs. Several papers have been published or submitted for publication based on the content of this thesis. Table 1.1 summarises these papers, showing their contributions and the corresponding sections of the thesis.

Paper	Thesis Section	Contribution
Y. Zhou, Y. Xia, H. Yang, T. Morstyn (2024), Strengthened and Faster Linear Approximation to Joint Chance Constraints with Wasserstein Ambiguity, Under review for <i>INFORMS Journal on Computing</i> . Available at arXiv [72]	Chapter 3 and Appendix A	C1
Y. Zhou, C. Essayeh, T. Morstyn (2024), Aggregated Feasible Active Power Region for Distributed Energy Resources with a Distributionally Robust Joint Probabilistic Guarantee, <i>IEEE Transactions on Power Systems</i> . Published [73]	Sections 4.1–4.4 of Chapter 4	C2
Y. Zhou, C. Essayeh, T. Morstyn (2024), A novel surrogate polytope method for day-ahead virtual power plant scheduling with joint probabilistic constraints, <i>Electric Power Systems Research</i> , 234, Special Issue for Power System Computation Conference 2024, Paris. Published [74]	Sections 4.5–4.6 of Chapter 4	C2
Y. Zhou, C. Essayeh, S. Darby, T. Morstyn (2024), Evaluating the social benefits and network costs of heat pumps as an energy crisis intervention, <i>iScience</i> , 27(2). Published [49]	Chapter 5 and Appendix B	C3
Y. Zhou, C. Essayeh, T. Morstyn (2024), Datasets of Great Britain primary substations integrated with household heating information, <i>Data in Brief</i> , 54. Published [75]	Appendix B	C3
Y. Zhou, Á. Paredes, C. Essayeh, T. Morstyn (2024), AI-focused HPC Data centres Can Provide More Power Grid Flexibility and at Lower Cost, In review for <i>Advances in Applied Energy</i> . Available at arXiv [76]	Chapter 6 and Appendix C	C4

Table 1.1: List of Publications Included in the Thesis

1.3 Thesis structure

This thesis is organised into seven chapters, structured to address the research questions and contributions outlined in Section 1.2.

Chapter 2 provides an extended literature review, offering in-depth discussions on the existing body of work related to each of the three research questions. This chapter identifies key gaps in the literature, which the thesis seeks to address.

Chapters 3–6 present the work done to achieve the thesis contributions. These chapters answer the three research questions and correspond to different contributions as detailed in Table 1.1. The content includes details of the proposed methods, experimental setups, case study results, and discussions.

Finally, Chapter 7 concludes the thesis by summarising the key findings and discussing potential directions for future research. This chapter reflects on the broader implications of the work and suggests how the methodologies and insights developed in this thesis can inform future advancements in power system operation and planning.

Literature Review

Fully harnessing the flexibility of the growing number of DERs is crucial for developing cost-effective and sustainable energy systems. However, as discussed in the previous chapter, challenges persist in both the technical operation and understanding the fuel-poverty and power-system impacts of DERs along the pathway to unlocking their flexibility. This chapter provides a more detailed analysis and discussion of existing work addressing these challenges, underscoring key gaps that motivate the research undertaken in this thesis. Specifically, the chapter is organised as follows:

- Section 2.1 reviews existing methods in power system scheduling, including those adapted to integrate DERs. This section provides a fundamental understanding of power system scheduling and optimisation, which serves as the foundation for the subsequent discussion on challenges related to managing DER uncertainty.
- Section 2.2 examines methods for managing uncertainty in both general DER scheduling and DER aggregation, with a particular focus on CC and its extensions. This section reviews classical and distributionally robust approaches to uncertainty management, identifying their computational challenges and limitations.
- Section 2.3 reviews existing work on understanding the impacts of DERs, which extend beyond the technical challenge explored in earlier sections. This section provides a systematic review of related literature on major types of DERs, including EVs, distributed RES, distributed ES, HPs, and data centres.
- Finally, Section 2.4 analyses the reviewed literature and identifies key gaps. These gaps closely align with the research questions of this thesis as outlined in Chapter 1, particularly those concerning the computational complexity of managing DER uncertainty, the under-recognised role of HPs in addressing fuel poverty, and the flexibility potential of AI-focused data centres.

2.1 Power system scheduling

As discussed in Section 1.1.1, power systems already employ a scheduling framework to optimise operation by leveraging system flexibility. Within this framework, ED and UC are two typical examples. Both ED and UC share the objective of minimising the overall operational cost of the system. ED is typically implemented closer to real-time operation and, as a result, often employs a more fine-grained modelling approach. In contrast, UC is conducted at least a day ahead and accounts for the binary commitment status of generators, making it a mixed-integer programming (MIP) problem. This MIP complexity arises from the physical characteristics of fossil-fuel generators, which require hours to start up and shut down [77].

In both ED and UC, network constraints are not central components, as is the case in the current GB day-ahead market clearing [14]. However, since network constraints must ultimately be satisfied, it is intuitively beneficial to incorporate them—typically in approximate forms—into ED and UC models. The most common approach involves the use of the DCPF approximation, which is widely adopted in academic research and power system simulation tools [21, 22, 78]. Existing studies have shown that DCPF provides acceptable accuracy under certain conditions: when line flows are not close to their thermal limits, voltage profiles remain sufficiently flat, and the resistance-to-reactance ratio is below 0.25 [79].

However, as discussed in Section 1.1.1, even with the DCPF approximation, power system scheduling remains computationally challenging due to the large problem scale—particularly in UC, which is a MIP and thus an NP-hard problem. To address this challenge, decomposition techniques such as Benders decomposition [80] and Lagrangian relaxation [81] are commonly applied to simplify UC problems by breaking them into smaller subproblems. For ED, the alternating direction method of multipliers (ADMM) is a representative decomposition approach—a main reason is that the convergence to optimality is guaranteed by the problem's convexity under the DCPF model [82].

However, there exists a broad set of practical scenarios where the DCPF approximation is insufficiently accurate [18]. In such cases, the full ACPF model becomes important, motivating research on incorporating ACPF constraints into ED and UC formulations. Since ACPF is nonlinear and non-convex, solving such models is complicated. To address this, researchers have applied outer approximation techniques [83], algebraic methods such as holomorphic embedding [84], and convex relaxation in the forms of semidefinite programming (SDP) or second-order cone programming (SOCP), which become exact under specific conditions [85]. More recently, data-driven methods have also emerged. For instance, graphical neural networks (GNN) have been employed to learn the solution mappings of AC optimal power flow (ACOPF) problems, showing promising performance [20].

2.1.1 Scheduling of distributed energy resources

The emerging DERs can, in principle, be incorporated within the existing ED and UC frameworks. However, their integration significantly increases problem complexity. Firstly, as DERs are connected at the distribution level, their integration necessitates the inclusion of distribution network constraints. Distribution networks typically have higher resistance-to-reactance ratios, which limits the accuracy of the DCPF approximation [86], thereby requiring more accurate network models, such as the full ACPF formulation. Secondly, the integration of potentially millions of DERs drastically increases the problem scale [9], further exacerbating computational intractability. Thirdly, as transmission and distribution networks are often operated by different entities, privacy preservation becomes a critical consideration for coordination algorithms.

To address these three challenges, distributed optimisation is widely employed in the context of DER scheduling and coordination [87]. In distributed optimisation, the overall problem is decomposed into smaller subproblems—similar to the methods discussed in Section 2.1—which are then solved in a decentralised manner by different entities. A representative use case is the TSO-DSO coordination [88, 89]. This framework not only enhances computational efficiency but also preserves the privacy of individual subsystems [87]. A range of decomposition methods have been explored for this purpose, including Benders decomposition and ADMM [87–90].

Note that there are also heuristic and data-driven approaches to DER scheduling. Heuristic algorithms aim to find good-quality solutions to complex optimisation problems in a more computationally efficient manner. Common heuristic algorithms include simulated annealing, evolutionary algorithms, and swarm intelligence techniques such as particle swarm optimisation [91]. Ref. [92] proposed a heuristic signalled particle swarm optimisation method for short-term DER scheduling, achieving a comparable optimal cost to an optimisation-based method but with a $35\times$ speedup in computation. Data-driven methods, on the other hand, learn the solution mappings of optimisation problems from historical data, enabling efficient real-time implementation (at the millisecond level) once trained offline. In [32], a recurrent neural network (RNN) was trained to learn the solution mapping of an EV fleet scheduling problem, achieving an optimality gap of less than 10% compared to an optimisation-based method with imperfect information, and delivering an $800\times$ speedup in real-time implementation. Reinforcement learning (RL) and multi-agent reinforcement learning (MARL) are also popular approaches for DER scheduling, in which a (neural network) policy is trained by interacting with the environment and maximising a reward function (typically non-differentiable). Ref. [93] demonstrated that MARL can achieve a cost outcome between that of a naïve baseline and a centralised optimisation method (with perfect information) when scheduling DERs in

a residential sector. However, a key limitation of heuristic and data-driven methods is the lack of theoretical guarantee on the optimality or feasibility of the obtained solutions, which is important for safety-critical power system applications. For this reason, the main focus of this thesis is on optimisation-based approaches that provide such theoretical guarantees.

2.1.2 Aggregation of distributed energy resources

Although distributed algorithms are promising, a major issue is the considerable communication burden caused by the iterative exchange of information required to achieve convergence [29]. As the problem scale increases, more iterations are required to achieve convergence and thus more communication burden will be placed [30].

DER aggregation offers an alternative and promising approach for scheduling numerous DERs in a computationally affordable manner, and tackles the three challenges in DER optimisation discussed in the previous section. The most representative method in DER aggregation is to calculate the aggregated feasible power region (AFPR), which is essentially the feasible operational region for the aggregated power of a group of DERs. The AFPR is advantageous as it provides a mathematically rigorous guarantee of optimality equivalent to the direct system-wide scheduling, provided that the AFPR is calculated exactly [30]. Furthermore, an AFPR transforms aggregated DERs into a set of operational constraints for their aggregated power; this set of constraints is analogous to those of conventional power system assets such as power plants. In certain cases, the AFPR can even take the same form of constraints associated with generators or batteries, giving rise to the concepts of virtual generators (VGs) or virtual batteries (VBs) [94, 95]. These characteristics enhance the integration of AFPRs into existing power system operational regimes and also improve their interpretability and transparency.

Calculating the AFPR is essentially a geometric projection problem [96]. The original feasible region for individual DERs exists in a high-dimensional space, where each dimension corresponds to a decision variable for the respective DER or a network state variable such as line flow or nodal voltage. The constraints of DERs and the network define the feasible region in this high-dimensional space. The AFPR is the projection of this high-dimensional feasible region onto the dimensions that represent the aggregated power of the DERs. In other words, for any aggregated power within the AFPR, there must exist a corresponding set of DER powers that satisfy all the DER constraints and network constraints [30]. Unfortunately, deriving the exact geometric projection is generally NP-hard [96], which calls for approximation methods, and existing approximation methods can be broadly classified into three categories: geometric methods [97, 98], random sampling (RS)-based methods [99], and optimisation-based methods [25, 95, 100–102]:

- Geometric methods typically start with Minkowski addition, which identifies the set of all possible summed power outputs of individual DERs. Geometric properties such as homothets and zonotopes were investigated to derive the approximated Minkowski addition result [97, 98]. While these methods are computationally efficient, they fail to account for the network constraints, which are critical for ensuring the power system safety and reliability.
- RS-based methods involve randomly generating operational statuses for individual DERs. The feasible aggregated power trajectories are then evaluated by analysing network power flow solutions [99]. Although conceptually straightforward, these methods may require a large number of samples to achieve a high approximation quality, leading to excessive computational burden.
- Optimisation-based methods approximate the AFPR by solving mathematical optimisation problems that explicitly incorporate both DER operational constraints and network constraints. The objective can be minimising the conservativeness of the approximation [25, 100, 101] or a metric to guide the update of the approximation in the next iteration [95, 102]. These methods are particularly suited for capturing the complexities of modern power systems and addressing the network safety requirements.

Overall, DER aggregation is a promising approach to address the computational scalability challenges in large-scale DER scheduling. Once the AFPR is computed, it can be seamlessly integrated into existing power system scheduling frameworks, resulting in a smaller problem scale compared to direct scheduling. Moreover, this integration does not introduce communication or convergence issues as in distributed optimisation. However, the main drawback lies in the computational difficulty of calculating the AFPR, which is NP-hard. As a result, approximation schemes—such as geometric methods, RS-based methods, and optimisation-based methods—are required, albeit at the cost of increased conservativeness or reduced guarantees on feasibility and safety.

2.2 Uncertainty in power system scheduling

As discussed in Section 1.1.2, most DERs are inherently characterised by significant uncertainties, which arise from their dependence on external and often unpredictable factors such as weather conditions and human behaviours. To ensure a reliable power system operation, it is therefore important to consider uncertainty under the increasing trend of DER integration, which is also the first research question of this thesis *Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering the DER uncertainty?*

2.2.1 Dealing with uncertainty in general optimisation problems

In an optimisation problem, uncertainty can appear in either the objective function or within the constraints. When uncertainty appears in the objective, it is typically addressed by minimising the expectation, which can be approximated using the sample mean (the sample average approximation). Illustrative examples can be found in [103].

This thesis focuses primarily on uncertainty in constraints, as power systems have high reliability requirement on the satisfaction of safety-critical constraints. When there is uncertainty in constraints, one approach is to ensure feasibility in all possible scenarios, known as the RO framework, or more specifically, the robust counterpart framework [104]. Analytical reformulations for RO have been developed for many practical optimisation problems [104], and its “worst-case” guarantee (conditioned on a given uncertainty set) is appealing to power system operators. However, RO often results in overly conservative solutions [39, 104], and it is less obvious to set the size of the uncertainty set. Chance-constrained programming (CCP) offers a promising alternative [105]. CCP supports robust but economic decision-making by explicitly limiting the probability of constraint violations, allowing for a controlled degree of conservativeness relative to RO. Also, compared to setting the size of the uncertainty site, it becomes more intuitive to set the constraint violation probability. As a result, CCP has found widespread applications in fields such as power systems, economics, finance, and water management [106].

Chance-constrained optimisation

A classical CCP can be defined as follows:

$$\min_{x \in \mathcal{X}} C(x) \quad (2.1a)$$

$$\text{s.t. } \mathbb{P}[\xi \notin \mathcal{S}(x)] \leq \epsilon, \quad (2.1b)$$

where $C(\cdot)$ is the objective function, $\mathcal{X} \subset \mathbb{R}^L$ is a compact domain for the decision variables $x \in \mathbb{R}^L$, and $\mathbb{P}$ is a probability measure associated with the random vector $\xi \subseteq \mathbb{R}^K$, characterizing its distribution. Here one can replace $\mathbb{P}$ with $\mathbb{P}_x$ to express a CCP with decision-dependent uncertainty (the distribution of ξ depends on the value of x). Constraint (2.1b) ensures that the random vector ξ falls outside the decision-dependent safety set $\mathcal{S}(x)$ with a small probability ϵ . Note that the form of (2.1b) follows recent studies in CC [107, 108], and can better highlight the fact that the probability function $\mathbb{P}(\cdot)$ is applied to the random vector ξ . The safety set can be expressed as:

$$\mathcal{S}(x) := \{\xi \mid f_p(x, \xi) \geq 0, p \in [P]\}, \quad (2.2)$$

where $[P]$ denotes the index set collecting all integers between one to the integer P (inclusively), namely $[P] := \{1, \dots, P\}$. When $P = 1$, the CC (2.1b) is an individual CC (ICC). When $P > 1$, the CC (2.1b) becomes a joint CC (JCC) [105], which requires all constraints to be met simultaneously with a high probability (note that there can be several ICCs as well as several JCCs in a single optimisation problem). Therefore, compared to ICC, JCC is a safer alternative under the same risk level ϵ , and is often more desired in practical applications [38, 106].

This thesis specifically focusses on CC with a linear form of $f_p(\mathbf{x}, \boldsymbol{\xi})$, due to its applicability in many practical cases and the availability of analytical reformulations [106, 107]. According to [107, 108], CC with linear $f_p(\mathbf{x}, \boldsymbol{\xi})$ can be further classified into two categories: CC with right-hand-side uncertainty (RHS-CC) and CC with left-hand-side uncertainty (LHS-CC). The RHS-CC has the following form of $f_p(\mathbf{x}, \boldsymbol{\xi})$:

$$\text{RHS: } f_p(\mathbf{x}, \boldsymbol{\xi}) = \mathbf{b}_p^\top \boldsymbol{\xi} - \mathbf{a}_p^\top \mathbf{x} + d_p, \quad (2.3)$$

where $\mathbf{a}_p$, $\mathbf{b}_p$, and d_p are known parameters. In comparison, the LHS-CC has the following form of $f_p(\mathbf{x}, \boldsymbol{\xi})$:

$$\text{LHS: } f_p(\mathbf{x}, \boldsymbol{\xi}) = (\mathbf{b}_p - \mathbf{A}_p^\top \mathbf{x})^\top \boldsymbol{\xi} - \mathbf{a}_p^\top \mathbf{x} + d_p, \quad (2.4)$$

where $\mathbf{A}_p$ is also a known parameter matrix. Compared to RHS-CC, LHS-CC has an extra product term between the decision variable $\mathbf{x}$ and the random vector $\boldsymbol{\xi}$.

As there is no such product coupling in RHS-CC, the computational tractability is greatly improved. Specifically, for a RHS-ICC, this constraint can be easily transformed into a deterministic form based on the inverse cumulative distribution function (CDF) $F_{\mathbf{b}_p^\top \boldsymbol{\xi}}^{-1}(\epsilon)$ with respect to the random variable $\mathbf{b}_p^\top \boldsymbol{\xi}$ at the risk level ϵ :

$$\mathbb{P}(\mathbf{a}_p^\top \mathbf{x} \leq \mathbf{b}_p^\top \boldsymbol{\xi} + d_p, p \in [1]) \geq 1 - \epsilon \quad (2.5a)$$

$$\Leftrightarrow \mathbf{a}_p^\top \mathbf{x} \leq F_{\mathbf{b}_p^\top \boldsymbol{\xi}}^{-1}(\epsilon) + d_p. \quad (2.5b)$$

However, if the joint constraint satisfaction is desired (RHS-JCC), the computational challenge appears again. RHS-JCC is in general non-convex. Convexity is preserved only under specific conditions, such as when the probability density function (PDF) of $\boldsymbol{\xi}$ is log-concave, as in the case of a Gaussian distribution [105]. Furthermore, even when convex, RHS-JCC generally does not have a closed-form transformation that can be handled by commercial optimisation solvers, and need to be solved using iterative schemes such as the supporting hyperplane method [74, 105].

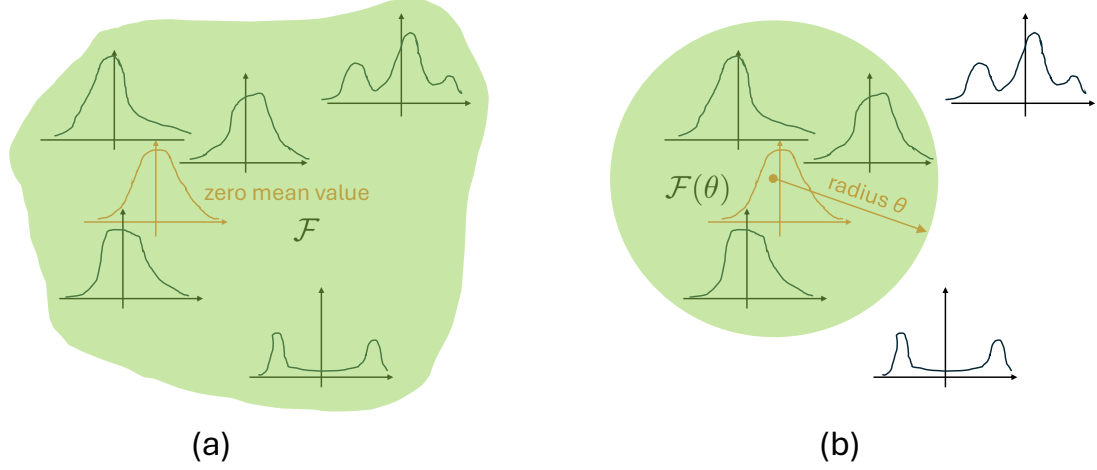


Figure 2.1: Illustration of ambiguity sets in distributionally robust optimisation. The green areas represent the ambiguity sets. Assuming all the distributions in this figure have zero mean values. (a) An example of a moment-based ambiguity set, which includes all distributions that share the same zero mean as the reference distribution. Some distributions with implausible shapes are included, which can lead to over-conservativeness. (b) An example of a distance-based ambiguity set, which includes distributions within a distance θ of the reference distribution. The two distributions with implausible shapes are excluded—even though they share the same mean as the reference distribution.

Alternatively, sample average approximation (SAA) can handle JCC with general distributions by first approximating the original distribution with a discrete empirical distribution constructed from historical data and then assigning uniform probability to each historical scenario [109]. However, this approach results in a MIP formulation with big-M parameters, which is NP-hard and is prone to numerical instability [106]. The computational challenge is exacerbated when more samples are used to have a higher approximation quality of SAA.

Another approximation solution method is the scenario-based approach [110], which replaces the original JCC with repeated constraints under different scenarios of random variables. This method is straightforward, requiring only the convexity of constraints $f_p(\mathbf{x}, \xi)$ and $\mathcal{X}$. It also provides a confidence level for the satisfaction of the original JCC based on the number of scenarios used in the reformulation. However, compared to SAA, the scenario-based method requires more restrictive conditions to recover the feasibility and optimality of the original CCP [106].

Finally, RO techniques can also be applied to solve CCP by constructing appropriate uncertainty sets that provide a safe approximation of the original CCP. A review of the RO-based solution schemes is provided in [106].

Distributionally robust chance-constrained optimisation

Classic CC models typically rely on the exact probability distribution of random variables. However, in many practical decision problems, including DER scheduling, decision-makers would only have access to a historical dataset. This dataset is often insufficient to accurately infer the true distribution of the random variables, limiting the out-of-sample performance of classical CC methods [111]. To address the ambiguity of the true distribution, DRCC has been proposed. DRCC aims to control the violation probability under the worst-case probability distribution within a so-called “ambiguity set” [111, 112]. The ambiguity set is a collection of probability distributions that are considered plausible given available historical data, and can be defined in various ways, such as the set of probability distributions sharing the same statistical moments (mean and variance) [113], or as distributions within a certain distance θ from a reference distribution derived from available historical data (namely distributions within a θ -radius ball centered at the reference distribution) [114]. Note that some studies also refer to the distance-based ambiguity sets as metric-based ambiguity sets [111]. Despite the good computational tractability, moment-based ambiguity sets do not fully leverage the information contained in the dataset. As illustrated in Fig. 2.1(a), these moment-based sets can include an overly broad range of distributions, including those with significantly different and implausible shapes, which can thus lead to overly conservative solutions [39, 114]. In contrast, as can be seen in Fig. 2.1(b), distance-based ambiguity sets can exclude the distributions with implausible shapes.

For distance-based ambiguity sets, common distance measures include the ϕ -divergence (also known as f -divergence) and the Wasserstein distance [39]. This thesis specifically focuses on the Wasserstein distance, a metric that quantifies the distance between two probability distributions based on optimal transport theory [39, 111, 114]. Compared to the ϕ -divergence, the Wasserstein distance is a more appropriate measure, as it satisfies the triangle inequality and symmetry properties [111, 114]. Moreover, the Wasserstein distance is more effective in capturing the shape of distributions [114] and can offer stronger out-of-sample performance [111].

In general, a problem with DRCC can be formulated as:

$$\min_{\mathbf{x} \in \mathcal{X}} c(\mathbf{x}) \quad (2.6a)$$

$$\text{s.t.} \quad \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{P}[\boldsymbol{\xi} \notin \mathcal{S}(\mathbf{x})] \leq \epsilon, \quad (2.6b)$$

Constraint (2.6b) ensures that the random vector $\boldsymbol{\xi} \subseteq \mathbb{R}^K$ falls outside the decision-dependent safety set $\mathcal{S}(\mathbf{x})$ with a small probability, which is not greater than the risk level ϵ under the worst-case distribution $\mathbb{P} \in \mathcal{F}$, with $\mathcal{F}$ being the ambiguity set specified by the decision maker.

Similar to the classical CCP case, handling a distributionally robust individual chance constraint (DRICC) is easier, and there are convex reformulation schemes for DRICC under moment-based ambiguity sets [115]. However, solving the joint version, known as distributionally robust joint chance constraint (DRJCC), remains computationally challenging. Tractable and convex reformulations are only possible under specific structural assumptions, such as the RHS structure for moment-based ambiguity sets, combined with additional requirements on the support and statistical moments [116]. For DRCC with the Wasserstein ambiguity set (WDRCC)—recognised for their favourable properties—only MIP reformulations are available for the more general WDRJCC with LHS structure [117], the special RHS structure [108, 118], and even the RHS-DRICC structure [108, 118]. It should be noted that MIP formulations benefit from advancements in commercial solvers, such as Gurobi [119], which are efficient for solving MIP problems. However, the NP-hard nature of these formulations limits their applications to large-scale problems. To tackle this challenge, iterative solution schemes have been proposed, including methods like ALSO-X, ALSO-X+, and ALSO-X#, which (approximately) solve problems with JCC or DRJCC iteratively with high-quality solutions [120, 121]. For better computational tractability, other approximation schemes were also investigated. Bonferroni approximation is a typical and tractable convex inner-approximation that approximates the original JCC by several ICCs with smaller risk levels, which, however, can become overly conservative when events overlap significantly [122]. The worst-case conditional value at risk (W-CVaR) method was also explored as a convex inner-approximation to WDRJCC [111, 122], which is also a general approximation scheme for CC. Ref. [122] proposed the “best” convex inner-approximation for RHS-WDRJCC, which is equivalent to W-CVaR under certain hyperparameters.

2.2.2 Chance-constrained power system scheduling

Many power system scheduling problems can use CC to achieve a reliable satisfaction of constraints (satisfaction with a high probability). There are three main categories of power system applications that involve active research in integrating CC. The first is the power system ED problem, the second is the power system UC problem, and the last category embed ED or UC in a multi-level optimisation setting where different levels may correspond to decision-makers with conflicting objectives. Although DERs were not intensively integrated to power systems when ED and UC were first proposed and applied, these two problems can naturally incorporate DERs, and this integration has been increasingly demonstrated in recent research, as will be reviewed below.

CC naturally arises in power system operations when certain constraints must be satisfied with a high probability. One typical example for CC is related to (distributed) wind farms. Dispatchable wind farms need to have power dispatch levels that are available with a high probability, and the total wind utilisation is typically required to exceed a specified threshold

with a high probability [123]. The increasing DER penetration further increases the importance of this CC formulation. Another example is on EVs, where CC is useful to address the uncertain availability and energy requirements of EVs. This modelling has been explored in a power system ED problem with EV integration [124]. Finally, when uncertainty arises from non-dispatchable sources or DERs, such as non-controllable loads or non-dispatchable (distributed) renewable generation, the uncertainty appears in power system network constraints, reserve requirements, and thermal generators that provide reserves (known as the participation factors of generators [106], automatic generation control (AGC) [125], or droop control coefficients [38]). In these cases, CC can be applied to ensure the high-probability satisfaction of these constraints [125, 126].

To solve CC-based ED or UC problems, solution schemes from general CCPs have been extensively applied, incorporating power system characteristics. Even for the simplest RHS-ICC in Eq. (2.5), consideration of power system characteristics can lead to improved solution quality. For example, wind generation and its forecasting errors have been observed to follow non-Gaussian distributions in several countries, including the US, Spain, and Sweden [127]. A solution tailored for Gaussian distribution may thus lead to reduced out-of-sample performance (reduced constraint satisfaction or increased operational cost). To tackle the non-Gaussian issue, existing work in ED [128, 129] has investigated more accurate modelling of random distributions, employing versatile distributions capable of fitting non-Gaussian data while admitting analytical forms for both the cumulative distribution function (CDF) and its inverse. However, these versatile distribution approaches do not exploit the correlation structure across uncertainty sources or across time steps. To overcome this limitation, subsequent research has employed Gaussian mixture models [126, 130] or beta kernel density estimation [131], which better capture the dependencies and improve modelling accuracy.

The aforementioned research [126, 128–131], which explored distributional improvements, followed the ICC regime. As discussed earlier, JCC is often more desirable for ensuring simultaneous satisfaction of multiple constraints. Solving problems with JCC is significantly more challenging and requires more restrictive assumptions. For instance, Ref. [132] exactly solved the UC problem with a JCC by assuming that the wind generation follows a Gaussian distribution. This assumption ensures that the corresponding RHS-JCC is convex, allowing it to be solved using the supporting hyperplane method discussed in Section 2.2.1.

Incorporating non-Gaussian properties remains crucial for realistic modelling, especially for widespread DERs that have been consistently observed to follow non-Gaussian distributions, such as wind generation. SAA methods [123, 133–136] and scenario-based approaches [137–139], which require fewer distributional assumptions, were applied to ED and UC problems. However, these methods are computationally challenging. Notably, the convexity requirement

of scenario-based approaches limits their applicability to UC problems that are inherently non-convex MIP. Ref. [140] addressed this limitation by proposing an improved scenario-based approach capable of handling the non-convexity of CC-based UC problems, thus extending the applicability of scenario-based methods to more complex cases.

Similar to what has been reviewed for the general CC methods in Section 2.2, power system researchers have also recognised the limitations of classical CC, particularly the potential bias introduced by the limited amount of historical data, which leads to degraded out-of-sample performance. This recognition has motivated the adoption of DRCC in power system applications. Various types of ambiguity sets have been explored for ED and UC problems with DRCC. Moment-based ambiguity sets, which rely on statistical moments such as mean and variance, have been investigated in [38, 141–143]. These sets provide tractable formulations under specific structural assumptions, making them a popular choice for integrating DRCC into power system problems. Distance-based ambiguity sets have also gained significant interest due to the better utilisation of historical data and the reduction of over-conservativeness, as has been studied in [125, 144–147]. Specifically, Wasserstein ambiguity set has attracted specific interest from power system researchers due to the superior out-of-sample performance [144–147]. A systematic review of DRCC applications in power systems can be found at [148].

Finally, in addition to standard UC or ED, there exists another form of complex power system multi-level optimisation, where ED or UC is embedded into the lower levels. These multi-level optimisations often model decision-makers with conflicting objectives at different levels, such as 1) strategic bidding problems for price-maker energy market participants, where the upper-level objective is the participant's profit maximisation and the lower level is market clearing (ED) [11, 149–151]; 2) merchant transmission investment, where the upper-level objective is the merchant transmission company's profit maximisation with a lower-level ED [152, 153]; and 3) sequential and coupled market clearing, such as unit commitment with gas market awareness [17, 154]. Even in the absence of conflicting objectives, a multi-level setting remains relevant when the optimal dual variables of a lower-level problem (such as energy market prices) need to be explicitly modelled [155]. As CC has been integrated into ED and UC to hedge the uncertainty introduced by the surging DER integration, it is natural that these power system multi-level problems will also incorporate CC-based ED in their lower-level

formulations, leading to:

$$\begin{aligned}
 & \max_{x \in \mathcal{X}} F(x, y) \\
 & \text{s.t. } y \in \arg \min_{y \in \mathcal{Y}} f(x, y) \\
 & \text{s.t. } \mathbb{P}[\xi \notin \mathcal{S}(x, y)] \geq 1 - \epsilon, \\
 & \quad z \in \arg \min_{z \in \mathcal{Z}} h(x, y, z) \\
 & \text{s.t. } \dots
 \end{aligned} \tag{2.7}$$

To solve these multi-level problems, optimality conditions such as the Karush–Kuhn–Tucker (KKT) conditions or strong duality are typically exploited to derive tractable single-level counterparts, which require convexity at the lower levels. As reviewed in Section 2.2.1, the non-convexity of most CC imposes a significant barrier to this class of problems. Consequently, existing work has either considered ICC [156] or applied grid-search algorithms to enumerate possible upper-level decisions [157] in order to hedge DER uncertainty in the lower-level problems.

2.2.3 Uncertainty in aggregation of distributed energy resources

As discussed in Section 2.1.2, the large-scale integration of DERs leads to large problem scale of power system scheduling, and thus a direct system-wide optimisation becomes computationally intractable, which motivates the hierarchical aggregation method for efficiently exploiting DER flexibility. Although aggregation overcomes the computational issue compared to direct scheduling, the challenge caused by managing DER uncertainty persists.

The integration of risk management is naturally aligned with RS-based approaches in DER aggregation. Building upon the RS methods in a deterministic setting, a risk-aware RS approach can be constructed by retaining samples that are feasible for a specific portion of realisations of DER uncertainty, as proposed in [158, 159]. However, RS-based approaches are already computationally intensive even in the deterministic case. The aforementioned risk-aware extension essentially repeats the RS process multiple times, further increasing computational complexity. Ref. [160] also employed an RS-based approach to hedge uncertainty. To reduce complexity, they limited the number of DER realisations to 25, justifying this choice based on an ergodicity argument—each scenario could represent a large number of possible realisations when there are numerous DERs to be aggregated. Nevertheless, 25 samples still introduce non-trivial additional computational complexity, and this justification lacks a theoretical guarantee. Overall, the limitations above may explain why most approaches to risk-averse DER aggregation are optimisation-based. However, it should also be noted that RS-based approaches have a unique advantage in handling non-convex DER models and network models, as demonstrated in [160].

A significant body of work on DER aggregation under uncertainty employs optimisation-based approaches, integrating RO or CC to hedge against uncertainty. Ref. [161] formulated the problem into a two-stage RO problem, ensuring that for any desired aggregated reactive power, and for any realisation of uncertainty, there always exist disaggregatable DER actions. Ref. [162, 163] extended this idea to address both active and reactive power aggregation. The aforementioned works [161–163] primarily focused on single-time-period aggregation. Ref. [164] advanced this line of research by applying the RO framework to derive the AFPR for multi-period aggregation. This approach ensures that for any realisation of uncertainty, disaggregatable DER actions for several future time steps always exist, accounting for network constraints and time-coupling DER constraints such as the storage energy limits.

A notable downside of RO is its potential for over-conservativeness [106, 165], which has motivated the exploration of CC methods for AFPR. ICC was applied to ensure compliance with the individual power limits of uncertain DERs [166], network constraints [94, 95], and the maximum capability of the aggregated DER power [94, 95]. However, CC methods rely on the assumption that the underlying distribution is known, which can be unrealistic as discussed in Section 2.2.1. To address this limitation, DRICC was applied to the aggregation process, and was used to enforce the maximum aggregated power limits [167], network voltage constraints [167, 168], the derivation of maximum-volume homothet approximations for individual DERs [169], and the maximum-volume approximation for the aggregated AFPR [170].

2.3 Impacts of distributed energy resources

The challenges in unlocking DER flexibility extend beyond the technical aspects of managing the computational complexity of DER scheduling with JCC. It is equally important to tackle challenges related to understanding the impacts of DERs that are being rapidly integrated, such as their impacts on power systems, economics, and even the society (e.g., poverty or equity). This lack of understanding can impede the strategic planning of DER rollout and may affect the social acceptance of DERs, which would then delay the DER adoption and thus the realisation of DER flexibility. Typical DERs include EVs, distributed RES, distributed ES, HPs, and data centres. On the demand side, as illustrated in Fig. 2.2, EVs, HPs, and data centres together are projected to account for up to 38% of the total electricity demand in GB as the country transitions towards net-zero carbon emissions. The remaining demand largely comprises other domestic electricity usage and industrial loads. On the generation side, distributed RES—those connected to the distribution networks—can contribute up to 20% of the total electricity generation in GB. The rest of the supply is made up of transmission-connected RES, nuclear power, and low-carbon gas turbines (equipped with carbon capture technologies). ES does not directly generate nor consumes energy and is thus not displayed

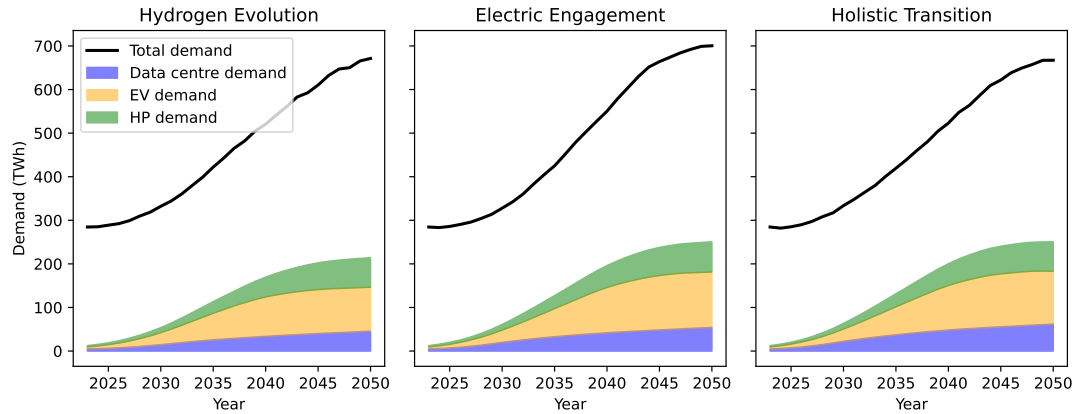


Figure 2.2: Future annual demand of typical DERs in GB.

“Electric Engagement”, “Hydrogen Evolution”, and “Holistic Transition” are three pathways that achieves net-zero emissions as defined in [9]. The y-axis represents annual electricity consumption. The data is sourced from UK FES 2024 [9].

in the figure; but based on UK FES 2024 [9], the total capacity (GWh) of distributed battery-based ES can be as high as 80% of the capacity of battery-based ES connected at the transmission level. Existing studies have contributed to evaluating the impacts of specific DERs across multiple dimensions, especially for EVs, distributed RES, and distributed ES.

2.3.1 Impact of electric vehicles

As a major emerging electricity load, the impacts of EVs have been studied since the early 1980s [171], and an extensive body of literature now exists examining their effects from multiple perspectives. Numerous studies have demonstrated the economic benefits of EVs, either directly or indirectly. For instance, [172] showed that a 5% increase in EV uptake could reduce distribution charges by approximately 2%. Similarly, [173] demonstrated that EV integration could lower wholesale electricity prices by 0.6–0.7% by 2050. Other studies have explored the use of EVs as home batteries during peak hours to enable household cost savings, which can outweigh battery degradation costs when the tariff differential is high or when a separate battery is purchased for use beyond a certain number of cycles [174].

In addition to economic impacts, the social implications of EVs have also been widely studied. For example, [175] found that rural households face greater financial barriers to EV adoption, while [176] and [177] identified disparities in adoption rates and projected cost parity for low-income households by 2029. Broader concerns around “transport poverty”—the inability to access essential transport services—have also been examined. For instance, [178] showed that EV car-sharing represents a particularly cost-effective approach to alleviating transport

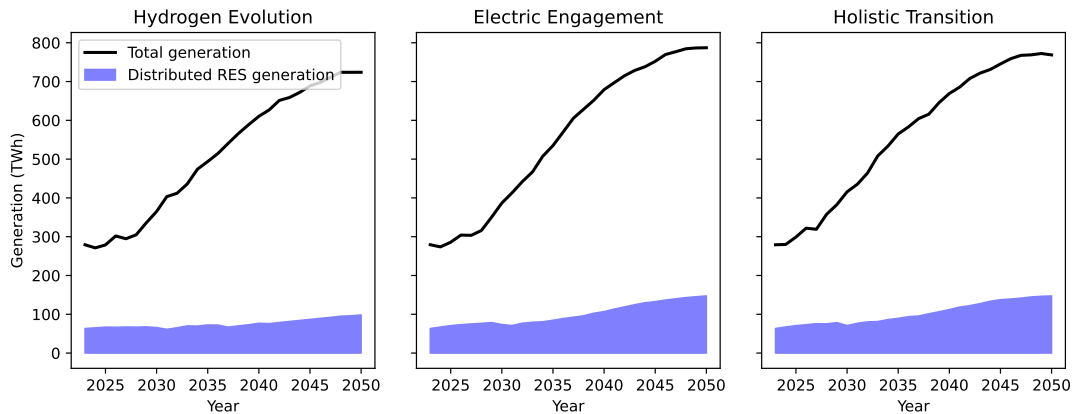


Figure 2.3: Future annual generation of RES connected to distribution networks in GB.

“Electric Engagement”, “Hydrogen Evolution”, and “Holistic Transition” are three pathways that achieves net-zero emissions as defined in [9]. The y-axis represents annual electricity generation. The data is sourced from UK FES 2024 [9].

poverty. Furthermore, [179] identified a “double vulnerability”, whereby households experience both energy poverty and transport poverty, with the two potentially reinforcing each other. They argued that policy for the rollout of new technologies should explicitly consider the interlinkages between these two dimensions of poverty.

Existing work has also examined the impacts of EVs on electricity demand and power networks. These impacts have been assessed across various national contexts, including the UK [180], the Netherlands [181], and the United States [182, 183]. At the distribution network level, studies have identified issues on transformer over-loading [184], voltage drops [185], and broader infrastructure challenges [186–188]. Such negative network impact can be mitigated by leveraging the power system flexibility of EVs, a concept commonly referred to as smart charging. For example, [189] demonstrated a 20% reduction in network upgrade costs through smart EV charging. Similarly, [8] found that EV flexibility could reduce the proportion of distribution network upgrades from 28% to 9%, using a Monte Carlo simulation to address uncertainty in EV charging demand.

Overall, existing studies have examined the economic, social, and power system impacts of EVs, highlighting their potential to reduce costs, alleviate transport poverty, and mitigate adverse network impact through smart charging.

2.3.2 Impact of distributed renewables

The literature on distributed RES is as extensive as that on EVs. Studies have demonstrated the economic benefits of distributed RES in enhancing system efficiency and financial viability [190], as well as the corresponding advantages under different policy regimes [191]. Ref. [192] further evaluated the coordination of distributed RES and storage (including EVs), forecasting that the levelised cost of a PV plus storage system will continue to decrease and fall below the grid electricity tariff in China. Similarly, Ref. [193] showed that integrating distributed RES into a hybrid energy system can achieve a carbon intensity up to 70% lower than that of the Australian grid (in 2015), while maintaining a levelised cost comparable to the grid electricity tariff.

The social impacts of RES have also been examined. For instance, [194] found that RES can help address the trade-off between income equality and environmental quality. When the share of RES in total final energy consumption exceeds approximately 39%, the marginal effect of income equality on reducing the ecological footprint becomes positive and statistically significant—meaning that both improved income equality and a smaller ecological footprint can be achieved simultaneously. This is because the low-carbon RES starts to meet the increased energy demand associated with higher income equality. Ref. [195] observed disparity issues in current PV deployment, where it was found that PV adoption occurs predominantly among affluent households, with certain incentive policies disproportionately benefiting wealthier households. They also proposed potential mitigation measures, including targeting subsidies to lower-income households, encouraging community solar facilities, and better publicizing the characteristics of the incentive programs, especially in vulnerable communities. Furthermore, [196–198] highlighted the potential of RES to alleviate fuel or energy poverty due to the higher efficiency and lower running cost of distributed RES. In particular, [196] estimated the elasticity of energy poverty with respect to distributed RES to be between -50% and -60% for European Union countries.

Extensive research has also focused on the power system impact of distributed RES. The well-known “duck curve” phenomenon—resulting from the mismatch between solar generation peaks and electricity demand peaks—has been analysed and reported in various studies [199, 200]. Other research has focused on the network impact such as reverse power flow and voltage rise [201, 202]. Broader reviews have also assessed RES impacts on power losses, voltage levels, frequency regulation, system balancing, and short-circuit current levels [203]. Similar to EV, these negative network impact can also be alleviated by exploiting flexibility, specifically on the real and reactive power control enabled by the RES inverter system. Existing work has demonstrated such benefits in improving the hosting capacity and reducing the overvoltage by 25–60% in Australia power grids [204, 205].

Overall, existing research on distributed RES has exploited their low running costs, low emissions, inverter-based flexibility, and projected reductions in upfront costs, based on which their impacts across multiple dimensions, including power systems, economics, and society, were studied.

2.3.3 Impact of distributed energy storage

There are also a range of studies examining the impact of distributed ES from various perspectives. Regarding economic benefits, Ref. [206] showed that using distributed ES for peak shaving can lead to a 1–39% reduction of the household's non-DR electricity bill under realistic tariffs. The operation of distributed ES can also affect the economic benefits, and [207] demonstrated that centralised operation of distributed home ES would be more cost-saving compared to decentralised operation. Ref. [208] compared the economic savings from distributed ES across different battery technologies, highlighting the superior cost-saving of lithium-ion batteries. Additionally, Ref. [209] found that in integration of distributed ES can lead to a 6.4–49.1% cost-saving for all participants in a peer-to-peer local electricity market.

From a social perspective, Ref. [210] highlighted the socioeconomic and demographic disparity issue in access to distributed ES in California. Ref. [211] observed that certain policy measures can simultaneously boost the adoption of distributed ES and advance energy equity. With regard to fuel poverty, most existing studies focus on combined PV and distributed ES systems, noting that the addition of ES enhances the fuel poverty reduction achieved through PV alone [212–214].

Existing work also evaluated the impact of distributed ES on the power demand and network. Unlike EV and RES, which may cause network issues under normal and uncontrolled operation, the impacts of distributed ES are generally positive, as it is typically deployed as an additional source of flexibility by default. A wide range of studies have investigated control strategies for realising the flexibility of distributed ES, which have been shown to reduce peak demand [215–217]. For example, [218] found that 2 kWh of battery storage per household could potentially halve peak demand at low-voltage substations. Regarding network impact, most studies report positive outcomes from integrating distributed ES, including improved transient stability [219], increased load and generation hosting capacity [220], reduced power losses and network unbalance [221], and mitigation of issues caused by distributed PV integration [222,223]. Moreover, distributed ES has been shown to offer greater voltage magnitude regulation compared to centralised ES [224].

Overall, existing research on distributed ES is as extensive as that on EVs and distributed RES, covering dimensions across power networks—predominantly with positive impacts—as well as economics and society.

2.3.4 Impact of heat pumps

Similar to EVs, distributed RES, and distributed ES, the economic impact of HPs was also investigated. As a key technology for heating electrification, HPs have gained significant attention. One of most well-known characteristics is their high running efficiency—typically around 300%. Existing studies have extensively examined the economic implications of this efficiency. Refs. [53, 225, 226] demonstrated that HPs can achieve a running cost comparable to gas boilers while offering additional carbon reduction benefits. However, the high upfront cost of HPs, which is several times greater than that of gas boilers, remains a critical barrier [53]. Ref. [227] illustrated the importance of the UK Renewable Heat Incentive (RHI) in offsetting the high upfront costs of HPs, making them more financially viable. Similarly, Ref. [228] showed that HPs can be competitive with gas boilers in Italy when reasonable refurbishment incentives are provided. However, this competitiveness may be compromised by recent increases in energy prices for both natural gas and electricity [229] (note that the UK gas and electricity prices are highly related as gas is still the main generation source, which was around 25% in 2024 [50]). In the UK, Ref. [230] identified air-source HPs as the cost-optimal low-carbon heating solution. Nevertheless, Ref. [52] warned that HPs can result in higher running costs compared to gas boilers, especially for cold regions.

However, the social impact of HPs remain less investigated compared to EVs, distributed RES, and distributed ES. Among these fewer studies, Ref. [231] showed that the locational pricing mechanism in electricity markets can lead to higher HP costs in regions with projected high heat demand, resulting in geographical disparities; Ref. [232] found that expansionary economic activities, such as those triggered by the production and installation of HPs, can help mitigate the contractionary pressures caused by rising energy prices; Ref. [52] showed that HPs can lead to slightly higher running costs compared to gas boilers, potentially increasing fuel poverty in the UK if gas boilers are replaced with HPs.

Widespread heating electrification can also increase electricity demand. Existing studies have examined the impact of HP uptake on peak electricity demand under various scenarios. In the GB context, Ref. [233] showed that the peak electricity demand in GB will increase by 14% if 20% of households adopt HPs; Ref. [234] estimated that the peak electricity demand will increase by 65% under the 100% household HP adoption, while another study projected this demand increase to 100% [235]. Refs. [236] and [237] analysed the economic benefits and the corresponding increase in electricity demand for various US cities.

As discussed in Section 1.1.3, peak demand alone is insufficient to understand the network impact. Existing works on HPs are aware of this insight and provide evaluation on the network impact directly. Ref. [238] found that HPs cause greater network constraint violations compared to EVs in Dutch low-voltage distribution grids, while EVs tend to induce more prolonged violations. Ref. [239, 240] analysed the distribution network impact of both HPs and EVs, offering a broader perspective on their combined effects. Ref. [241] advanced the discussion

by quantifying the network upgrade cost required for HP deployment. Similarly, Ref. [242] found that HPs can lead to significant network upgrade costs—up to approximately £1,000 per dwelling for rural grids, which is notably higher than the upgrade cost for urban grids in Belgium. Ref. [243] made a similar observation that the rural grids may have more issues, and overloading and voltage issues will begin to emerge when HP penetration reaches 30% in Belgian rural distribution feeders. Ref. [244] further analysed the impact of HPs on distribution networks in Northern California. These network issues call for effective strategies to unlock the HP flexibility.

2.3.5 Impact of data centres on power system flexibility

In parallel to EVs and HPs, data centres represent another significant source of electricity demand, primarily driven by computing workloads [245]. The power consumption of data centres has been observed to be correlated with server utilisation [246], which is closely linked to the execution of computing jobs. This inherent relationship provides data centres with the flexibility to provide power system services, enabling them to adjust their power consumption by flexibly scheduling computing jobs. This characteristic also drives a huge portion of existing work focused on the amount of new power system flexibility that can be unlocked from data centres. This is also the main focus of this thesis.

Applications of flexible data centres

In fact, there have been a wide range of attempts to exploit data centre flexibility for various applications that are beyond power system applications. Data centre flexibility in job scheduling can be leveraged to optimise data centre-specific metrics, such as maximising computing throughput. These applications together with power system applications share a common fundamental framework centred on the scheduling of computing jobs. This correlation facilitates the transfer of methods and insights across different application domains. Therefore, the review of data centre flexibility in this section is not limited to power system applications.

Many real-world large-scale data centres have exploited job flexibility in their operations. For example, certain jobs can be executed on different machines, low-priority jobs can be delayed or preempted, and job scheduling can be adjusted; these operations have shown to improve data centre utilisation, reduce queue delays, or optimise other metrics concerned by data centre operators [247]. Several scheduling systems have been deployed for this purpose. Google developed Borg [247], a data centre scheduler that identifies machines capable of meeting a job's resource requirements and employs a scoring model to minimise stranded resources. Microsoft introduced Tetris [248], which matches multi-resource task requirements with machine resource availability, thereby increasing cluster efficiency. Similarly, Alibaba proposed the Fuxi scheduling system [249], achieving 95% CPU utilisation by flexible job

scheduling. Given the complex requirements, interdependencies, and scheduling controllability in data centres, job scheduling is an NP-hard problem [250]. Consequently, learning-based approaches have also been explored as a computationally tractable way for flexible data centre scheduling [251–253].

In addition to improving data centre utilisation and speeding up computations, a significant body of work focuses on enhancing energy efficiency. For instance, Refs. [254–256] optimised data centre energy efficiency by strategies such as allocating tasks to more energy-efficient servers or minimising uninterruptible power supply (UPS) losses, accounting for varying energy efficiency levels based on device utilisation. Decarbonisation has also emerged as a key area of interest. Refs. [250, 257–261] explored methods for flexibly scheduling jobs during periods with lower grid emission rates, thereby reducing the carbon footprint of computing operations. This concept has been successfully demonstrated by Google in their real-world data centre operations in 2021 [262]. Google effectively limited hourly capacity when the grid's energy supply mix was carbon-intensive and delayed the execution of temporally flexible workloads to align with periods of “greener” energy supply.

Various existing work has also assessed data centre flexibility in providing power system services [11]. For example, Refs. [69, 70] investigated the role of data centre flexibility in maintaining power system frequency. In Ref. [263], data centre flexibility was applied to mitigate voltage and imbalance issues in distribution networks. Ref. [264] explored the use of data centre flexibility for regulation services, where a real-world data centre successfully tracked power system signals every few seconds. Additionally, Ref. [265] examined how data centre flexibility could participate in energy balancing markets. These studies collectively demonstrated the potential of data centre flexibility to contribute to a wide range of power system services, highlighting its versatility and importance in supporting grid stability and reliability.

Data centre quality of services

An important consideration in data centre flexible scheduling is the quality of service (QoS), as low QoS can lead to reduced user satisfaction and retention, directly impacting the data centre's profitability and long-term viability. To account for QoS, much of the existing literature focuses on limiting job delays by imposing deadlines for job completion or simultaneously minimising delays as an extra objective. Examples of such works include [260, 261, 266–269]. This concept is often referred to as Service Level Agreements (SLA) [260, 268, 269]. However, a key limitation of these approaches is that the delays are not monetised. As a result, it remains unclear whether it is sensible to use data centre flexibility for applications that may compromise QoS, such as power system applications.

Recently, a few studies began to explicitly incorporate the QoS cost into their optimisation objectives [250, 270]. These studies modelled the QoS cost based on the compensation schemes outlined in the SLA of mainstream cloud computing platforms, such as Alibaba [271], Google Cloud [272], Amazon AWS [273], and Oracle [274], where users receive credits when QoS is degraded.

AI-focused data centres

As mentioned in Section 1.1.3, the rapid growth of AI applications has led to the emergence of AI-focused data centres. These AI-focused data centres are characterised by their heavy reliance on energy-intensive GPUs, which are essential for training and running AI models. Consequently, AI-focused HPC data centres tend to be more energy-demanding than traditional general-purpose CPU-intensive data centres. EPRI estimates that data centres now consume approximately 4% of total power generation in the United States, and this proportion could increase to as much as 9.1% by 2030, with AI being one of the main drivers [64]. Recently, technology companies have also announced plans for hyperscale AI data centres with significant power consumption. Meta has announced plans to build a 2 GW data centre in Manhattan [275], and xAI has reported a similar plan for its own facility, which is intended to house 1 million AI GPUs with a total power consumption also around 2 GW [276]. These developments highlight the growing importance of AI-focused data centres in the energy landscape. This surge in power consumption has placed considerable stress on power system capacity, leading power system operators to delay new connection requests from data centres. An industry survey [277] found that 44% of data centre respondents reported average quoted utility wait times exceeding four years. In this situation, many data centre operators (60% of respondents [277]) are considering deploying on-site power generation systems (such as fossil-fuel generators) as their top-ranked option, which raise significant environmental concerns [278].

Recently, the rapid growth of AI-focused HPC data centres is attracting increasing attention from researchers. Ref. [279] observed that AI-focused data centres require schedulers tailored to the operational characteristics of GPUs, in contrast to traditional schedulers designed primarily for CPUs. To address this need, improved scheduling strategies have been proposed, leveraging job flexibility to achieve various operational objectives. For example, several studies have aimed to reduce average queuing and execution times [280–283], while others have focused on improving cost efficiency, including reductions in power consumption and financial expenditure for cloud service rental [71, 284–286]. In addition, some works have targeted the reduction of carbon emissions in AI-focused data centres by exploiting job flexi-

ility [287, 288]. Ref. [289] identified transient power quality issues during LLM training, which may cause instability issue in the power grid. Despite these advancements, there remains a notable gap in the literature concerning the unique characteristics of AI-focused data centres in the context of providing power system flexibility services.

2.4 Gap analysis

So far, this chapter has reviewed literature related to the key challenges in unlocking the grid flexibility of DERs. This body of literature spans topics from managing the uncertainty of DERs to understanding their impacts, such as impacts on power systems, economics, and society. However, there are still several key research gaps which motivate the research questions of this thesis.

2.4.1 Gaps in managing the uncertainty of distributed energy resources through chance constraints

The first set of research gaps relate to the technical challenge of unlocking DER flexibility (the first research question), namely, utilising DER flexibility while managing DER uncertainty through CC. Gaps are present in both standard CC-based DER scheduling and CC-based DER aggregation.

Gap analysis: chance-constrained power system scheduling

In CC-based DER scheduling, as has been reviewed in Section 2.2.2, it is evident that there is significant value in and a clear need for DRCC and specifically DRJCC due to its joint safety guarantees and higher reliability compared to classical JCC. Specifically, this thesis focuses on WDRJCC, given its strong out-of-sample performance as discussed in Section 2.2.1 and the particular interest in recent power system applications [144–147]. However, the real-world power system application of WDRJCC remains challenging due to the excessive computational complexity of WDRJCC and the intrinsic complexity of power system scheduling problems [290]. As reviewed in Section 2.2.1, solving WDRJCC involves either MIP-based exact reformulations or iterative methods. The exact MIP-based reformulations of WDRJCC are NP-hard, while iterative schemes are more prone to numerical instability. Greater computational tractability remains a key challenge, particularly given that the original deterministic power system scheduling problems are already computationally intensive, such as the complicated UC problem with complexity exacerbated by DERs, or multi-level optimisation that requires lower-level convexity but also require lower-level CC to hedge DER uncertainty, as has been discussed in Section 2.2.2. Although convex approximation methods have been proposed that

have better tractability, these approximations each have particular limitations. The Bonferroni approximation [122] can become overly conservative when events overlap significantly. W-CVaR [111, 122] and the “best” convex inner-approximation [122] introduce a large number of additional ancillary variables and constraints, which bring excessive computing burden.

Overall, there remains a significant need for more tractable and, specifically, convex WDRJCC reformulations that meet the computational requirements in practical applications. Addressing this gap, Chapter 3 of this thesis will introduce a strengthened and faster linear approximation (SFLA) for RHS-WDRJCC, designed to improve computational efficiency while preserving solution quality. The development of such a method paves the way for scalable and practical implementations of chance-constrained power system scheduling problems, addressing critical challenges in modern power system operations.

Gap analysis: dealing with uncertainty in aggregation

As mentioned in Sections 2.2.1 and 2.2.2, JCC is often preferred over ICC. However, in the context of AFPR, there is a lack of work utilising JCC due to its inherent intractability and non-convexity that is against by many approximation methods for deriving AFPR. Consequently, the aforementioned studies on reliable AFPR [94, 95, 166–170] employed ICC or DRICC. However, ICC applied under the same violation rate as JCC lead to a less reliable AFPR, as the joint violation rate is not guaranteed. Therefore, there is a pressing need for scalable methods that incorporate JCC while maintaining a balance between reliability and computational efficiency in deriving the AFPR. Chapter 4 of this thesis will address this gap by proposing a method that can integrate JCC or WDRJCC in a computationally tractable manner.

2.4.2 Gaps in the impact of distributed energy resources

The second class of research gaps lies in addressing challenges in understanding the impact of DERs. As discussed in Section 1.1.3, beyond the computational challenge, existing work has investigated the power-grid, economic and social impacts of many DERs. These studies also represent a crucial dimension that supports the design of effective and efficient roadmaps to realise DER flexibility. However, HPs and data centres are still under-investigated, which are major types of DERs as illustrated in Fig. 2.2. Given the pivotal role that HPs and data centres are expected to play in future power systems, this knowledge gap poses a significant challenge in unlocking the full value of DER flexibility.

Gap analysis: impacts of heat pumps on fuel poverty and distribution networks

Unlike the well-understood spectrum of social and economic impacts associated with EVs, distributed RES, and distributed ES, a key research gap remains for HPs in its under-investigated impact on fuel poverty. While Ref. [52] has assessed the fuel poverty impact of replacing all gas heating with HPs, no studies have evaluated the fuel poverty impact of replacing off-gas heating with HPs. In fact, the benefits of HPs are more on replacing off-gas heating instead of gas heating: existing research has extensively demonstrated that HPs are only comparable to gas boilers in terms of running costs [53, 225, 226, 228–230] and may result in higher costs [52]; coupled with their high upfront cost [53], HPs can be a less economically attractive option for homes connected to the mains gas grid unless proper government incentives are placed [227, 228]. However, as has been discussed in Section 1.1.3, HPs are particularly attractive to off-gas homes. These homes rely on expensive off-gas heating and are associated with the highest risk of fuel poverty [43, 54]. In 2023, there are still around 40% of UK homes using non-gas heating as their main heating methods over the winter [56] (this number is higher than the 15.7% off-gas-grid proportion because a large portion of households used portable heaters for that winter [56], possibly due to the high energy bills). Therefore, there is an urgent need to understand the potential value of HPs in reducing energy costs particularly for these off-gas homes, which help address the critical fuel poverty issue.

The second key gap lies in understanding the associated network upgrade cost after replacing off-gas heating with HPs. As reviewed in Section 2.3.4, existing studies have illustrated that distribution networks may require substantial upgrades under various HP uptake scenarios [238–244]. This underscores the importance of evaluating regional network impact when HPs are used to replace all off-gas heating, as these costs represent a critical consideration for policymakers. Despite the importance of this issue, no existing studies have specifically examined replacing off-gas heating with HPs as a distinct scenario in their network analysis.

Chapter 5 of this thesis will address these two gaps, by carrying out a detailed evaluation of the fuel poverty and network impact of HPs if they were to replace off-gas heating based on household datasets for the 32 local authorities in Scotland, 9 government regions, and network data for around 4,000 primary substations (in distribution networks) in the GB.

Gap analysis: impacts of data centres on power system flexibility

Although Section 2.3.5 shows that a large amount of existing work has evaluated the potential of power system flexibility for data centres, we identified three important gaps in this area. Specifically, as has been pointed out in Section 2.3.5, the first gap in the existing literature is the lack of studies that consider the distinct characteristics of AI-focused data centres in the context of power system flexibility services. AI-focused HPC data centres are more power-demanding due to their heavy reliance on GPUs, which may result in different capabilities for providing power system services compared to traditional general-purpose HPC data centres

that rely on CPUs. Additionally, recent work [71] has identified distinct job patterns between the two types of data centres, further influencing their potential to support power system operations. Understanding these differences is crucial for both data centre operators and power system operators, because this knowledge can motivate enhanced collaboration to develop data centre flexibility schemes tailored to the specific characteristics of different types of data centres.

The second gap lies in the modelling of the cost of data centre flexibility. While incorporating QoS constraints (see [260, 261, 266–269]) provides some insights, existing work failed to link QoS impacts to a monetary cost of flexibility. Therefore, there remains a lack of incentives for data centre operators to provide flexibility, especially to provide power system services that most data centre operators are unfamiliar with. Although a few recent work [250, 270] quantified the QoS cost based on SLA, most cloud platforms—such as Google, Amazon, and Alibaba—use availability metrics rather than time-of-delay metrics [271–274]. In these availability metrics, credits are assigned based on resource unavailability rather than job delays. However, providing grid flexibility does not necessarily require shutting down servers and can be achieved through hardware power regulation techniques like power capping and dynamic voltage and frequency scaling (DVFS) [291–293]. Consequently, the true cost of providing power system flexibility that will not shut down servers but only delay computation remains unclear. Overall, there is a lack of a unified framework to model the data centre flexibility cost for different grid flexibility services, highlighting a critical gap that requires further investigation.

The last gap is the lack of simultaneous assessment of data centre flexibility for various power system services. Most existing research focused on data centre flexibility for specific services [69, 70, 263–265], without providing a unified evaluation of how data centres can contribute to a broad range of power system services. These fragmented evaluations makes it challenging to evaluate and compare the overall potential and limitations of data centre flexibility and to identify which grid services are best suited for what types of data centres. A holistic assessment is important to better understand the role of data centres in supporting various grid services and to guide the effective integration of data centre flexibility into power system operations.

Chapter 6 of the thesis will address these gaps by analysing and comparing the flexibility potential of AI-focused and general-purpose HPC data centres across a wide range of power system services. Furthermore, Chapter 6 will propose a method to quantify the cost of flexibility provision using real-world computing price data, offering valuable insights for both data centre operators and power system operators.

Strengthened and Faster Linear Approximation of Wasserstein Chance Constraints

This chapter and the next chapter, Chapter 4, answer the first research question of this thesis:

Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering the DER uncertainty?

As discussed in Section 1.1.2 and Section 2.2.1, uncertainty arises frequently in power system operations especially under the soaring penetration of DERs. WDRJCC represents a promising method that 1) guarantees the simultaneous satisfaction of multiple constraints in DER scheduling and 2) improves the out-of-sample performance compared to classical JCC when the true distribution of the random variables is unknown. This chapter specifically focuses on developing a Strengthened and Faster Linear Approximation (SFLA) to WDRJCC with RHS uncertainty (RHS-WDRJCC), and demonstrates the superiority performance on two typical complicated real-world power system applications with DER integration, including the UC problem and the bilevel strategic bidding with a chance-constrained lower-level ED problem, as discussed in Sections 2.2.2 and 2.4.1.

For convenience, the definition of an optimisation problem with WDRJCC from Chapter 2 is replicated below:

$$\min_{\mathbf{x} \in \mathcal{X}} c(\mathbf{x}) \tag{3.1a}$$

$$\text{s.t.} \quad \sup_{\mathbb{P} \in \mathcal{F}_N(\theta)} \mathbb{P}[\xi \notin \mathcal{S}(\mathbf{x})] \leq \epsilon, \tag{3.1b}$$

where $c(\cdot)$ is the objective function, $\mathcal{X} \subset \mathbb{R}^L$ is a compact domain for the decision variables $\mathbf{x} \in \mathbb{R}^L$. Constraint (3.1b) ensures that the random vector $\xi \in \mathbb{R}^K$ falls outside the decision-dependent safety set $\mathcal{S}(\mathbf{x})$ with a small probability, which is not greater than the risk level ϵ under the worst-case distribution $\mathbb{P} \in \mathcal{F}_N(\theta)$. Here, the Wasserstein ambiguity set $\mathcal{F}_N(\theta)$ is defined as a ball with radius θ centered at the empirical distribution $\mathbb{P}_N$ constructed by

3. Strengthened and Faster Linear Approximation of Wasserstein Chance Constraints 41

historical data:

$$\mathcal{F}_N(\theta) := \{\mathbb{P} \mid d_W(\mathbb{P}_N, \mathbb{P}) \leq \theta\}. \quad (3.2)$$

Following [294] and [108], this chapter considers the 1-Wasserstein distance based on a general norm $\|\cdot\|$:

$$d_W(\mathbb{P}, \mathbb{P}') := \inf_{\Pi \in \mathcal{P}(\mathbb{P}, \mathbb{P}')} \mathbb{E}_{(\xi, \xi') \sim \Pi} [\|\xi - \xi'\|], \quad (3.3)$$

where $\mathcal{P}(\mathbb{P}, \mathbb{P}')$ is a set of distributions with marginal distributions $\mathbb{P}$ and $\mathbb{P}'$. RHS-WDRJCC gets its name because of the special RHS structure of the safety set as below:

$$\mathcal{S}(\mathbf{x}) := \{\xi \mid \mathbf{a}_p^\top \mathbf{x} \leq \mathbf{b}_p^\top \xi + d_p, p \in [P]\}, \quad (3.4)$$

where there are P constraints indexed by $[P] := \{1, \dots, P\}$ that need to be met jointly with high probability $1 - \epsilon$ with $\epsilon \in (0, 1)$. This joint satisfaction is desired in practical applications for its higher safety [38, 106], and the RHS structure also arises in several practical problems, such as managing the thermal constraints of the power grid lines, securing the power system reserve [126, 295, 296], or restricting the load-generation imbalance [132].

RHS-WDRJCC (3.1) is non-convex, and can be exactly reformulated as a set of MIP constraints [108, 294] with big-M parameters, which is NP-hard and can be computationally demanding. As has been discussed in Section 2.4.1, there are also iterative solution schemes, and convex approximations. However, these methods are either not applicable, computationally expensive, or overly-conservative, especially for the power system UC and the strategic bilevel bidding problems with a lower-level chance-constrained ED.

One promising solution approach is utilising valid inequalities, which are additional constraints introduced to an optimisation problem to tighten the formulation without excluding any feasible solutions to the original problem. Motivated by the work of [108] where valid inequalities were exploited to strengthen the exact MIP reformulation so as to speed up computation, this chapter shows that the strengthening process can also be applied to a convex inner-approximation that admits W-CVaR equivalence. This migration reduces the number of constraints in the original convex approximation and tightens the feasible region for ancillary variables, which leads to significant computational speedups. It should be noted that migration of a strengthening process can introduce higher conservativeness to a convex inner-approximation, as shown in Case I of Figure 3.1. However, this chapter proves that the proposed SFLA will always lead to Case II in Figure 3.1, which does not introduce additional conservativeness and can even produce a less conservative inner-approximation. Extensive numerical case studies further demonstrate the computational superiority of the proposed SFLA. In the UC problem, the proposed SFLA achieves up to 10× and on average 3.8× computational speedup compared to the existing strengthened and exact MIP reformulation [108]

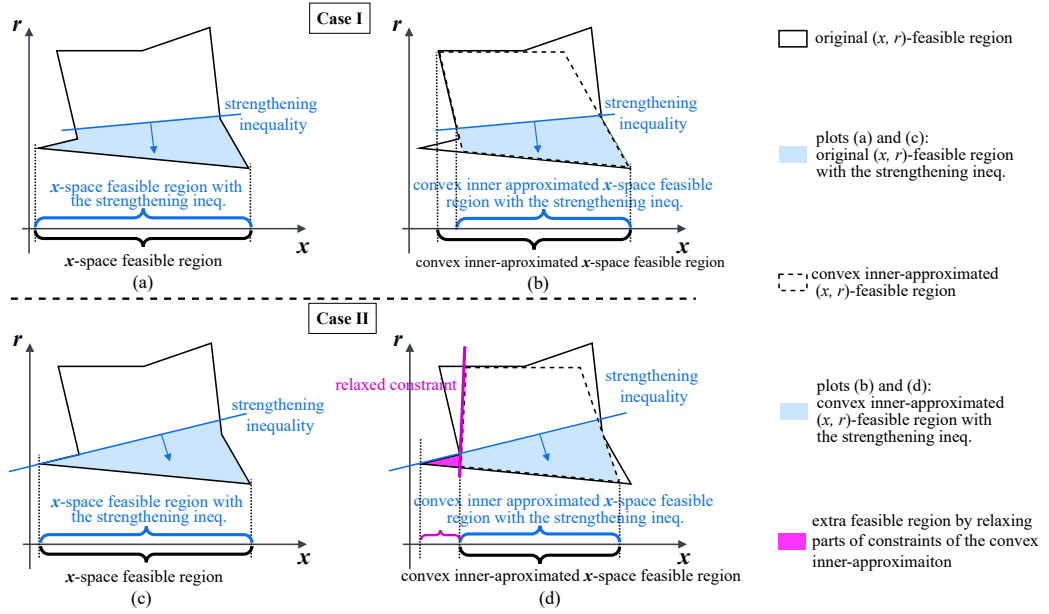


Figure 3.1: Two possible cases when a strengthening inequality valid for the original feasible region is applied to the convex inner-approximation.

Analytical reformulations sometimes (such as for WDRJCC) introduce ancillary variable r in addition to the original decision variable x . Case I: A strengthening inequality is valid for the original nonconvex region but leads to higher conservativeness to its inner-approximation. (a) In the original nonconvex feasible region of x and r , the strengthening (valid) inequality excludes a redundant portion without reducing the feasible region of x . (b) The convex inner-approximation (dashed) has a smaller feasible region of x compared to the original region, and the strengthening inequality further reduces the x feasible region, resulting in even higher conservativeness. Case II: A strengthening inequality is valid for both the original nonconvex region and its inner-approximation. (c) In the original nonconvex feasible region, the strengthening inequality excludes a redundant portion without reducing the feasible region of x . (d) The convex inner-approximation (dashed) has a smaller feasible region of x compared to the original region, and the strengthening inequality does not further reduce the x feasible region. Furthermore, by relaxing certain constraints of the convex inner-approximation, it is possible to include an extra non-empty feasible region (depicted in purple) and thus expand the feasible region of x (less conservativeness). This extra feasible region is small relative to the excluded region; therefore, the overall search space is still reduced, preserving the computational efficiency provided by the strengthening inequality.

to find comparable high-quality solutions. In the bilevel strategic bidding problem where the exact formulation is not applicable due to non-convexity, the proposed SFLA demonstrates 90 $\times$ speedup in computation time compared to convex inner-approximations such as W-CVaR.

This chapter is organised as follows. Section 3.1 lists the existing exact reformulation and convex inner-approximations. Section 3.2 introduces the proposed SFLA and proves its relationship to exact reformulation and other convex approximations. Section 3.3 conducts numerical studies, and Section 3.4 concludes this chapter.

3.1 Problem formulation

This chapter focuses on RHS-WDRJCC (3.1b), with the safety set described in Eq. (3.4). In what follows, this chapter assumes $\epsilon \in (0, 1)$ and $\theta > 0$, which are conditions necessary for the subsequent exact reformulation on which this work is based [294]. This section summarises the current exact reformulation and approximation methods for RHS-WDRJCC.

3.1.1 Exact reformulation

Suppose N independent and identically distributed (i.i.d.) samples $\{\xi_i\}_{i \in [N]}$ are collected for the random vector ξ , where the index set is defined as $[N] := \{1, \dots, N\}$. Given a decision $x \in X$ and a sample ξ_i , this chapter defines the distance from ξ_i to the complement of $S(x)$ as:

$$\text{dist}(\xi_i, S(x)) = \inf_{\xi' \in \mathbb{R}^K} \{\|\xi_i - \xi'\| \mid \xi' \notin S(x)\}. \quad (3.5)$$

Building on the results given by [294] and the safety set defined in (3.4), $\text{dist}(\xi_i, S(x))$ can be reformulated as the following analytical expression:

$$\text{dist}(\xi_i, S(x)) = \left(\min_{p \in [P]} \frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top x}{\|\mathbf{b}_p\|_*} \right)^+ = \min_{p \in [P]} \left(\frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top x}{\|\mathbf{b}_p\|_*} \right)^+, \quad (3.6)$$

where $\|\cdot\|_*$ is the dual norm and $(\cdot)^+$ extracts the non-negative component of the argument, setting negative values to zero. By introducing ancillary variables $s \in \mathbb{R}$ and $\mathbf{r} \in \mathbb{R}^N$, [294] showed that RHS-WDRJCC (3.1b) can be expressed as the following distance-based formulation:

$$s \geq 0, \mathbf{r} \geq \mathbf{0}, \quad (3.7a)$$

$$\epsilon N s - \sum_{i=1}^N r_i \geq \theta N, \quad (3.7b)$$

$$\left(\frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top x}{\|\mathbf{b}_p\|_*} \right)^+ \geq s - r_i, \quad \forall i \in [N], p \in [P]. \quad (3.7c)$$

While this distance-based formulation cannot be directly processed by commercial solvers such as Gurobi [119], it can be transformed into an equivalent MIP formulation:

$$\mathbf{z} \in \{0, 1\}^N, s \geq 0, \mathbf{r} \geq \mathbf{0}, \quad (3.8a)$$

$$\epsilon N s - \sum_{i=1}^N r_i \geq \theta N, \quad (3.8b)$$

$$M(1 - z_i) \geq s - r_i, \quad \forall i \in [N], \quad (3.8c)$$

$$\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} + M z_i \geq s - r_i, \quad \forall i \in [N], p \in [P]. \quad (3.8d)$$

This MIP reformulation, based on [294], is originally designed for open safety sets, but [108] demonstrated that it applies to both open and closed sets.

Despite being solvable by advanced MIP solvers, MIP reformulation remains NP-hard and its complexity becomes particularly problematic when the original deterministic problem is already complicated or requires convexity for tractable reformulation such as in multilevel problems. Therefore, this chapter focuses on convex inner-approximation (for safety) of the $\mathbf{x}$ -feasible region $\mathcal{X}_{\text{Exact}}$, which is defined by the RHS-WDRJCC in (3.1) for the original decision variable $\mathbf{x}$:

$$\mathcal{X}_{\text{Exact}} := \{\mathbf{x} \in \mathcal{X} \mid \exists s, \mathbf{r} : (3.7a)-(3.7c)\}, \quad (3.9)$$

because this is the region that affects the objective value $c(\mathbf{x})$ in (3.1a).

3.1.2 Convex inner-approximations

This section introduces three representative convex inner-approximation methods for RHS-WDRJCC. The first is the Bonferroni approximation, which is a straightforward and computationally efficient inner-approximation that replaces a JCC/DRJCC with ICCs/DRICCs at reduced individual risk levels ϵ_p . According to [122], the Bonferroni approximation of RHS-WDRJCC (3.1b) can be formulated as:

$$\mathbf{a}_p^\top \mathbf{x} \leq d_p - \sup_{\mathbb{P} \in \mathcal{F}(\theta)} \mathbb{P}\text{-VaR}_{\epsilon_p}(-\mathbf{b}_p^\top \tilde{\boldsymbol{\xi}}), \quad \forall p \in [P], \quad (3.10)$$

where the preset individual risk levels ϵ_p must satisfy $\sum_{p \in P} \epsilon_p \leq \epsilon$, and the term

$$\sup_{\mathbb{P} \in \mathcal{F}(\theta)} \mathbb{P}\text{-VaR}_{\epsilon_p}(-\mathbf{b}_p^\top \tilde{\boldsymbol{\xi}})$$

is the optimal value of the following bilinear optimisation problem:

$$\min_{\alpha^B, \beta^B, \mathbf{m}, \eta} \eta \quad (3.11a)$$

$$\text{s.t.} \quad \theta\beta + \frac{1}{N} \sum_{i=1}^N \alpha_i^B \leq \epsilon_p, \quad (3.11b)$$

$$\alpha_i^B \geq 1 - m_i(\eta + \mathbf{b}_p^\top \boldsymbol{\xi}_i), \quad \forall i \in [N], \quad (3.11c)$$

$$\beta^B \geq m_i \|\mathbf{b}_p\|_*, \quad \forall i \in [N], \quad (3.11d)$$

$$\alpha^B \geq 0, \mathbf{m} \geq 0, \quad (3.11e)$$

which captures the risk of constraint violation under distributional uncertainty within a θ -radius Wasserstein ball and can be solved by a line search along η or bilinear solvers. Problem (3.11) is solved first, and the optimal value, namely $\sup_{\mathbb{P} \in \mathcal{F}(\theta)} \mathbb{P}\text{-VaR}_{\epsilon_p}(-\mathbf{b}_p^\top \tilde{\boldsymbol{\xi}})$, serves as parameters for the Bonferroni approximation (3.10).

The Bonferroni approximation (3.10) does not add extra variables or constraints to the original deterministic problem, resulting in minimal computational overhead. A common practice is to set $\epsilon_p = \epsilon/P$ [106], but this often leads to excessive conservativeness, especially when P is large or when the safety constraints are correlated [297]. However, even with optimised ϵ_p , the approach can still be overly conservative in some cases, and finding optimal values for ϵ_p is generally intractable [122]. As will be shown in the case studies, the Bonferroni approximation often leads to infeasibility due to this conservativeness. Consequently, this chapter excludes this approach from theoretical comparisons in the following discussion.

An alternative and widely applied inner-approximation is W-CVaR [111, 122], which can be formulated as the following set of linear constraints:

$$\alpha^W \geq 0, \beta^W \in \mathbb{R}, \tau \in \mathbb{R}, \quad (3.12a)$$

$$\tau + \frac{1}{\epsilon} \left(\theta\beta^W + \frac{1}{N} \sum_{i \in [N]} \alpha_i^W \right) \leq 0, \quad (3.12b)$$

$$\alpha_i^W \geq w_p \left(\mathbf{a}_p^\top \mathbf{x} - \mathbf{b}_p^\top \boldsymbol{\xi}_i - d_p \right) - \tau, \quad \forall i \in [N], p \in [P], \quad (3.12c)$$

$$\beta^W \geq w_p \|\mathbf{b}_p\|_*, \quad \forall p \in [P], \quad (3.12d)$$

where $\mathbf{w} := [w_p]_{p \in [P]}$ subject to $\mathbf{w} \in \Delta_{++} := \{\mathbf{w} \in (0,1)^P : \sum_{p \in [P]} w_p = 1\}$ is a tunable hyperparameter that affects the performance of the W-CVaR approximation by prioritising specific constraints in $\mathcal{S}(\mathbf{x})$ [122, 145]. Similarly, its $\mathbf{x}$ -feasible region can be defined as:

$$\mathcal{X}_{\text{WCVaR}}(\mathbf{w}) := \{\mathbf{x} \in \mathcal{X} \mid \exists \alpha^W, \beta^W, \tau : (3.12a)-(3.12d)\}. \quad (3.13)$$

Another convex approximation [122] reformulates the exact RHS-WDRJCC (3.7) by overcoming the non-convexity caused by the distance function (3.6). More specifically, this approximation method replaces the original distance function $\text{dist}(\xi_i, S(x))$ with a conservative approximation $\widehat{\text{dist}}(\xi_i, S(x))$ given by:

$$\widehat{\text{dist}}(\xi_i, S(x)) := \kappa_i \left(\min_{p \in [P]} \frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right), \quad (3.14)$$

where $\kappa_i \in [0, 1]$ are pre-selected parameters. Since $\widehat{\text{dist}}(\xi_i, S(x)) \leq \text{dist}(\xi_i, S(x))$, the replacement yields the following linear inner-approximation (LA) of the constraints in (3.7):

$$s \geq 0, \mathbf{r} \geq \mathbf{0}, \quad (3.15a)$$

$$\epsilon N s - \sum_{i \in [N]} r_i \geq \theta N, \quad (3.15b)$$

$$\kappa_i \left(\frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right) \geq s - r_i, \quad \forall i \in [N], p \in [P]. \quad (3.15c)$$

Ref. [122] showed that LA (3.15) represents a family of the “best” inner-approximation in the $(\mathbf{x}, s, \mathbf{r})$ -feasible region of (3.7) and LA (3.15) becomes exact by optimising $\boldsymbol{\kappa} := [\kappa_i]_{i \in [N]}$. Furthermore, as demonstrated by [122], LA (3.15) with $\boldsymbol{\kappa} = \mathbf{1}$ is equivalent to W-CVaR for a specific setting of $\mathbf{w}$, and becomes exact under certain conditions. These conditions are also extended to the proposed SFLA, as will be illustrated in Corollaries 4 and 5. The $\mathbf{x}$ -feasible region of the LA approximation can be expressed as:

$$\mathcal{X}_{\text{LA}}(\boldsymbol{\kappa}) = \{\mathbf{x} \in \mathcal{X} \mid \exists s, \mathbf{r} : (3.15a)-(3.15c)\}. \quad (3.16)$$

3.2 Strengthened and faster linear approximation

The proposed approximation is motivated by the strengthening of the exact MIP reformulation. As discussed, the RHS-WDRJCC reformulation (3.7) can be exactly reformulated as a set of MIP constraints in (3.8). [108] showed that the MIP (3.8) can be further strengthened to another exact MIP reformulation (denoted by ExactS), by replacing $NP - \lfloor \epsilon N \rfloor P$ constraints with P valid inequalities. For convenience, the formulation of ExactS is reproduced as Eq. (A.1) in Appendix A.1. The $\mathbf{x}$ -feasible region of ExactS remains equivalent to the exact MIP set (3.8), but the number of constraints and the complexity of the whole $(\mathbf{x}, s, \mathbf{r})$ -feasible region are reduced. This section will show that this strengthening procedure can be migrated to the RHS-WDRJCC (3.7) expressed by the distance function (Proposition 1), which then enables derivation of the proposed approximation, termed as *Strengthened and Faster Linear Approximation* (SFLA). Finally, rather than compromising the $\mathbf{x}$ -feasible region as illustrated in Case I

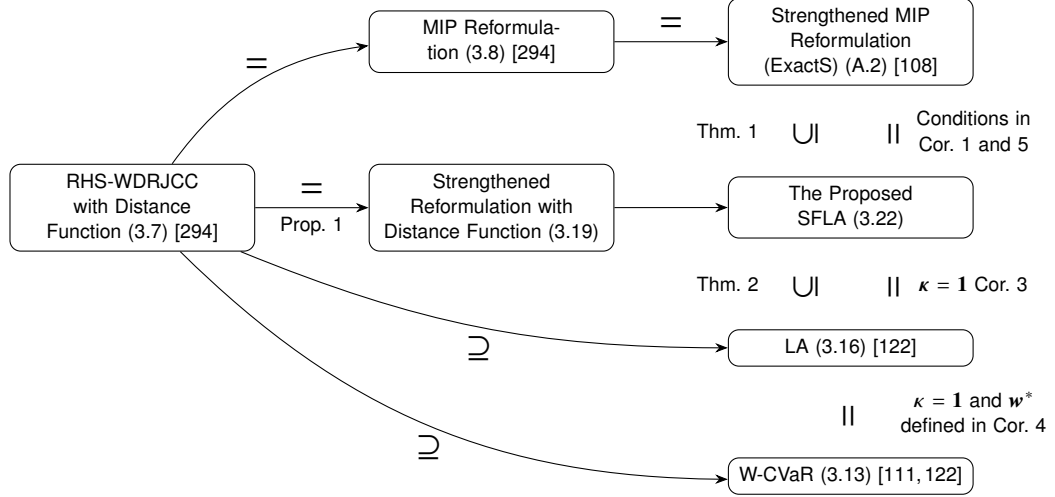


Figure 3.2: The derivation flow and the formulation comparisons of the x -feasible region defined by different reformulations of RHS-WDRJCC.

of Figure 3.1, this chapter proves that the proposed SFLA is always the Case II: the proposed SFLA will not introduce extra conservativeness and can even be less conservative than LA proposed by [122] (Theorem 2). Figure 3.2 displays the derivation process and a comparison of different formulations in the x -feasible region.

3.2.1 Derivation and formulation

Let $k := \lfloor \epsilon N \rfloor$ and assume an ordering $\mathbf{b}_p^\top \boldsymbol{\xi}_1 \leq \dots \leq \mathbf{b}_p^\top \boldsymbol{\xi}_N$ without loss of generality. This chapter further denotes by q_p the $(k+1)$ -th smallest value as:

$$q_p := \mathbf{b}_p^\top \boldsymbol{\xi}_{k+1}. \quad (3.17)$$

Then, this chapter introduces the index set $[N]_p$ with only $k = \lfloor \epsilon N \rfloor$ elements as:

$$[N]_p := \{i \in [N] \mid \mathbf{b}_p^\top \boldsymbol{\xi}_i < q_p\}, \quad \forall p \in [P]. \quad (3.18)$$

In addition, the following lemma is introduced:

Lemma 1 (Lemma 1 in [108]). *For any fixed $\mathbf{x} \in \mathcal{X}_{Exact}$, there exists $(\mathbf{r}, s)$ such that s is equal to the $(k+1)$ -th smallest value amongst $\{\text{dist}(\boldsymbol{\xi}_i, \mathcal{S}(\mathbf{x}))\}_{i \in [N]}$ and constraints (3.7a)–(3.7c) are satisfied.*

With Lemma 1, this chapter proposes the following proposition, which demonstrates that the strengthening procedures for the exact MIP reformulation (3.8) in [108] can be migrated to the exact RHS-WDRJCC reformulation (3.7) expressed by the distance function $\text{dist}(\boldsymbol{\xi}_i, \mathcal{S}(\mathbf{x}))$.

Proposition 1. *The feasible set $\mathcal{X}_{Exact}$ (3.9) can be equivalently defined by the following set of strengthened constraints, by replacing $NP - \lfloor \epsilon N \rfloor P$ constraints in (3.7c) involving the distance function with P valid inequalities (3.19d) and only keeping $\lfloor \epsilon N \rfloor P$ constraints in (3.7c) as (3.19c):*

$$s \geq 0, \mathbf{r} \geq \mathbf{0}, \quad (3.19a)$$

$$\epsilon N s - \sum_{i \in [N]} r_i \geq \theta N, \quad (3.19b)$$

$$\left(\frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right)^+ \geq s - r_i, \quad \forall i \in [N]_p, p \in [P], \quad (3.19c)$$

$$\frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq s, \quad \forall p \in [P]. \quad (3.19d)$$

In other words:

$$\mathcal{X}_{Exact} = \{\mathbf{x} \in \mathcal{X} \mid \exists s, \mathbf{r} : (3.19a)-(3.19d)\}. \quad (3.20)$$

The proof of Proposition 1 is provided in Appendix A.2. The form of constraints in (3.19) motivates us to propose SFLA, by replacing the non-convex constraints (3.19c) with the approximated distance function $\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))$ introduced in Eq. (3.14), while keeping the strengthening structure in (3.19). Therefore, the proposed SFLA can be expressed as:

$$s \geq 0, \mathbf{r} \geq \mathbf{0}, \quad (3.21a)$$

$$\epsilon N s - \sum_{i \in [N]} r_i \geq \theta N, \quad (3.21b)$$

$$\kappa_i \left(\frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right) \geq s - r_i, \quad \forall i \in [N]_p, p \in [P], \quad (3.21c)$$

$$\frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq s, \quad \forall p \in [P]. \quad (3.21d)$$

where there are still the preset hyperparameters $\kappa_i \in [0, 1]$ for all $i \in [N]$. The $\mathbf{x}$ -feasible region of the proposed SFLA can be expressed as:

$$\mathcal{X}_{SFLA}(\boldsymbol{\kappa}) = \{\mathbf{x} \in \mathcal{X} \mid \exists s, \mathbf{r} : (3.21a)-(3.21d)\}. \quad (3.22)$$

Comparing to the LA formulation (3.15), the proposed SFLA includes extra valid inequalities (3.21d) to tighten the solution space and brings computational speedup. At the same time, the $[N]$ set in LA (3.15c) is replaced with the $[N]_p$ set in SFLA (3.21c), which reduces the number of constraints from NP to $\lfloor \epsilon N \rfloor P$. This reduction is particularly beneficial when N is large, as it significantly decreases the number of constraints to be processed by solvers. The following section analyses the properties of $\mathcal{X}_{SFLA}(\boldsymbol{\kappa})$ under various choices of $\boldsymbol{\kappa}$.

3.2.2 Theoretical properties and analysis

This section demonstrates that the proposed SFLA set, $\mathcal{X}_{\text{SFLA}}(\kappa)$, forms an inner-approximation to $\mathcal{X}_{\text{Exact}}$ and can offer a less-conservative alternative to the LA set $\mathcal{X}_{\text{LA}}(\kappa)$. All the corresponding proofs can be found in Appendix A.2.

This section first shows that the proposed SFLA achieves an inner-approximation to RHS-WDRJCC:

Theorem 1. $\mathcal{X}_{\text{SFLA}}(\kappa)$ is a subset of $\mathcal{X}_{\text{Exact}}$, i.e., $\mathcal{X}_{\text{SFLA}}(\kappa) \subseteq \mathcal{X}_{\text{Exact}}$.

Observing Eq. (3.21c) in SFLA $\mathcal{X}_{\text{SFLA}}(\kappa)$, it can be seen that only $\lfloor \epsilon N \rfloor P$ constraints are replaced with the conservative approximated distance function $\widehat{\text{dist}}(\xi_i, \mathcal{S}(x))$. This selective replacement is central to making $\mathcal{X}_{\text{SFLA}}(\kappa)$ potentially less conservative than $\mathcal{X}_{\text{LA}}(\kappa)$. To further investigate this property, this chapter starts with demonstrating that $\mathcal{X}_{\text{LA}}(\kappa)$ admits a CVaR interpretation:

Lemma 2. $\mathcal{X}_{\text{LA}}(\kappa)$ in (3.16) is equivalent to the following set:

$$\left\{ x \in \mathcal{X} : \theta / \epsilon + \mathbb{P}_N\text{-CVaR}_{1-\epsilon}(-\widehat{\text{dist}}(\xi, \mathcal{S}(x))) \leq 0 \right\}, \quad (3.23)$$

where

$$\mathbb{P}_N\text{-CVaR}_{1-\epsilon}(v(\xi)) := \min_{s'} \left\{ s' + \frac{1}{\epsilon N} \sum_{i \in [N]} \max\{0, v(\xi_i) - s'\} \right\} \quad (3.24)$$

is the CVaR of the random variable $v(\xi)$ under the empirical probability distribution constructed by historical data $\mathbb{P}_N$ [105].

Based on Lemma 2, an extended version of Lemma 1 can be derived for the approximated distance function $\widehat{\text{dist}}(\cdot)$:

Lemma 3. For any $x \in \mathcal{X}_{\text{LA}}(\kappa)$, there exists (r, s) such that s is equal to the $(k+1)$ -th smallest value amongst $\{\widehat{\text{dist}}(\xi_i, \mathcal{S}(x))\}_{i \in [N]}$ and the constraints in (3.16), namely (3.15a)–(3.15c), are satisfied.

With all results above, this chapter now shows that the proposed SFLA $\mathcal{X}_{\text{SFLA}}(\kappa)$ is a less conservative (or at least equally conservative) approximation compared to the LA set $\mathcal{X}_{\text{LA}}(\kappa)$.

Theorem 2. The SFLA set $\mathcal{X}_{\text{SFLA}}(\kappa)$ is a superset of $\mathcal{X}_{\text{LA}}(\kappa)$, i.e., $\mathcal{X}_{\text{LA}}(\kappa) \subseteq \mathcal{X}_{\text{SFLA}}(\kappa)$.

Theorems 1 and 2 establish that the SFLA set achieves an approximation less conservative than LA and is an inner-approximation to RHS-WDRJCC as well, i.e., $\mathcal{X}_{\text{Exact}} \supseteq \mathcal{X}_{\text{SFLA}}(\kappa) \supseteq \mathcal{X}_{\text{LA}}(\kappa)$. Moreover, all properties of $\mathcal{X}_{\text{LA}}(\kappa)$ demonstrated by [122] naturally extend to the proposed $\mathcal{X}_{\text{SFLA}}(\kappa)$. For example, [122] showed (Proposition 8) that the good approximation

quality of $\mathcal{X}_{LA}(\kappa)$ is supported by the fact that its constraints (3.15a)–(3.15c) define the best set of inner convex approximations in the (x, s, r) -feasible region of the exact RHS-WDRJCC reformulation defined by (3.7). Therefore, the good approximation quality also holds for the proposed SFLA given Theorem 2.

In addition, this chapter shows that the SFLA can be exact by integrating κ as extra decision variables:

Corollary 1. *Let $\kappa \in [0, 1]^N$ be extra decision variables and define the updated SFLA set as:*

$$\mathcal{X}_{SFLA}^* = \{x \in \mathcal{X} \mid \exists \kappa \in [0, 1]^N, s, r : (3.21a)–(3.21d)\}.$$

Then it follows $\mathcal{X}_{SFLA}^* = \mathcal{X}_{Exact}$.

Taking $\kappa = \{\kappa_i\}_{i \in [N]}$ as extra decision variables renders the problem bilinear and thus difficult to solve. However, if κ is restricted to be uniform, namely $\kappa_i = \kappa$ or equivalently $\kappa = \kappa \mathbf{1}$ (recall that $\kappa_i \in [0, 1]$), then $\kappa = \mathbf{1}$ is the optimal parameter setting:

Corollary 2. $\mathcal{X}_{SFLA}(\mathbf{1}) \supseteq \mathcal{X}_{SFLA}(\kappa \mathbf{1})$ for $\forall \kappa \in [0, 1]$.

The proof is similar to that for Proposition 10 in [122]. In fact, for $\kappa = \mathbf{1}$, the proposed SFLA becomes equivalent to the approximation proposed by [122]:

Corollary 3. $\mathcal{X}_{SFLA}(\mathbf{1}) = \mathcal{X}_{LA}(\mathbf{1})$.

Note that, for $\kappa \neq \mathbf{1}$, it is possible that the proposed $\mathcal{X}_{SFLA}(\kappa)$ strictly dominates $\mathcal{X}_{LA}(\kappa)$, namely $\mathcal{X}_{SFLA}(\kappa) \supset \mathcal{X}_{LA}(\kappa)$. This argument has been numerically verified by comparing $\mathcal{X}_{SFLA}(\kappa)$ and $\mathcal{X}_{LA}(\kappa)$ under randomly generated κ subject to $\kappa \neq \mathbf{1}$.

In addition, the proposed SFLA $\mathcal{X}_{SFLA}(\kappa)$ becomes equivalent to W-CVaR in (3.12) with a specific setting of w and $\kappa = \mathbf{1}$:

Corollary 4. $\mathcal{X}_{SFLA}(\mathbf{1}) = \mathcal{X}_{WCVaR}(w^*)$, where $w^* = [w_p^*]_{p \in [P]}$ and $w_p^* = \frac{\|b_p\|_*^{-1}}{\sum_{l \in [P]} \|b_l\|_*^{-1}}$.

Finally, there are conditions when the proposed SFLA $\mathcal{X}_{SFLA}(\mathbf{1})$ becomes exact for $\kappa = \mathbf{1}$:

Corollary 5. $\mathcal{X}_{SFLA}(\mathbf{1}) = \mathcal{X}_{Exact}$ holds for either the following conditions:

1. $\xi_i \in \mathcal{S}(x)$ for all $i \in [N]$ and $x \in \mathcal{X}_{Exact}$.
2. $\epsilon \leq 1/N$.

The second condition in Corollary 5 shows that the proposed SFLA becomes exact when the number of data N is small or the risk level ϵ is low. This also implies that a smaller N and a smaller ϵ can lead to better approximation quality of the proposed SFLA. This implication is particularly important, as DRCCP is often used in cases with limited historical data due to its strong out-of-sample performance. Additionally, safety-critical applications, such as in power systems, generally demand a low risk level. Thus, the proposed SFLA can achieve high approximation quality across a range of practical applications.

Finally, as demonstrated by [294], a larger Wasserstein radius θ would convexify $\mathcal{X}_{\text{Exact}}$, which also suggests an increasing approximation quality of the proposed SFLA $\mathcal{X}_{\text{SFLA}}(\kappa)$ with the increase of θ .

3.3 Numerical experiments

This section presents numerical case studies to demonstrate the reduction of computational complexity and also the approximation quality of the proposed SFLA. Specifically, Section 3.3.1 investigates a single-level UC problem and Section 3.3.2 explores a bilevel storage strategic bidding problem. These two problems are typical complicated power system optimisation problems under the DER integration, as has been discussed in Section 2.2.2. For all norm calculations in the RHS-WDRJCC formulations, this section adopts the L2 norm which has the property of being self-dual.

Both of the numerical cases are programmed by Python 3.9.19 and solved by Gurobi 11.0.3 on a server with dual AMD EPYC 7532 CPUs and 256 GB RAM. The solver is configured as FeasibilityTol=1e-9, OptimalityTol=1e-9, IntFeasTol=1e-9, MIPGap=0.1%, a one-hour wall-clock time limit TimeLimit=3600s, and a maximum CPU thread number to be Threads=4; other solver parameters are set to default values. This section examines two risk levels $\epsilon \in \{0.025, 0.05\}$. For the Wasserstein radius θ , this section evaluates a small value that leads to a slightly less reliable result (the out-of-sample JCC satisfaction rate being lower than desired) and a large value that includes a slightly conservative result. This section sets $\kappa = 1$ for the proposed SFLA and LA, $w_p = 1/P$ for W-CVaR following [145], and $\epsilon_p = \epsilon/P$ for Bonferroni approximation. Detailed case study settings for the UC and bilevel problems, including datasets and the generation of random runs to ensure robust conclusions, are given in Appendices A.3.1 and A.4.1. The code alongside the mathematical formulations for the UC and bilevel problems are available at the following GitHub repository <https://github.com/PSAL0xford/SFLA>.

3.3.1 Unit commitment problem

The first case study examines a UC problem, where CCs are introduced to guarantee the system reliability in presence of DER uncertainty, specifically the uncertain wind generation. UC represents a typical real-world large-scale problem, and its primary objective is to determine the generator on/off statuses for the next scheduling period with a minimal operation cost [21]. UC is typically formulated as an MIP, where each generator requires at least T binary variables to represent on/off statuses over the next T time steps. With thousands of generators (e.g., 1400 in MISO), non-linear generator cost functions [77], and complex

network constraints, UC is challenging to solve even in the traditional power system operation and under a deterministic case. However, the soaring DER penetration introduces additional problem scale and increased uncertainty into UC, necessitating a computationally tractable approach to integrate JCCs.

This section specifically focuses on the comparison between the proposed SFLA and ExactS [108] in Eq. (A.1). Comparisons with LA [122] in Eq. (3.16), W-CVaR [111, 122] in Eq. (3.13), and Bonferroni approximation [122] in Eq. (3.10) are excluded here due to their significant disadvantages relative to the proposed SFLA, and the corresponding results are placed in Appendix A.3.2.

Evaluation metrics

The UC problem is primarily evaluated using the following metrics:

1. *TimeF* (s): This metric records the time (in seconds) to obtain the first “comparable high-quality” solution, defined as the solution with an objective within the MIPGap of the final objective achieved by the proposed SFLA (whether optimal or the best found within the *TimeLimit*). If SFLA does not yield a feasible solution within the *TimeLimit*, the *TimeF* of the proposed SFLA is set to the *TimeLimit*, and the *TimeF* for other benchmarks is set to the time when the first feasible solution is found (if any, otherwise set to the *TimeLimit*). Because MIP admits early stop, this metric offers a practical perspective in cases where the solver identifies a solution close to optimal early but requires extra time for verifying optimality.
2. *Time* (s): This measures the total time (in seconds) spent to solve the problem to optimality (within MIPGap), and is set to the *TimeLimit* if no optimal solution is found within the *TimeLimit*. This metric compliments the previous *TimeF* by providing verified optimality.
3. *Obj. Diff.* (%): This is the percentage difference between the optimal value (minimum UC cost) of an arbitrary benchmark method and that of the SFLA, relative to the optimal value of the SFLA. A positive value indicates that the benchmark achieves a lower cost (better optimality) than the SFLA. This chapter only keeps certain random runs where the benchmark method and the proposed SFLA are both solved to optimality.
4. *Reli.* (%): This is the out-of-sample joint satisfaction rate of the RHS-WDRJCC. This metric is used to justify the reasonability of the selection of Wasserstein radius θ , and is calculated on 5000 held-out samples.

This section prioritises *TimeF* over the total computation time *Time* due to comparison with ExactS. MIP can always have early-stopping and ExactS can have better optimality compared to inner-approximations (SFLA, LA, W-CVaR and Bonferroni). Therefore, a higher *Time* for ExactS does not necessarily suggest its disadvantage.

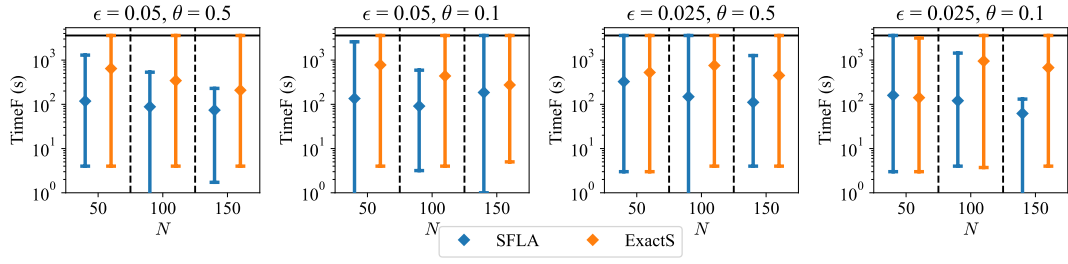


Figure 3.3: Computation time to obtain the first comparable high-quality solution for the proposed SFLA and the benchmark ExactS for the UC problem.

Dots represent the mean value of the 150 random runs, with error bars indicating the 95% percentile interval (from 2.5th to 97.5th). The black horizontal line represents the 3600s TimeLimit. The optimisation horizon is set to $T = 24$.

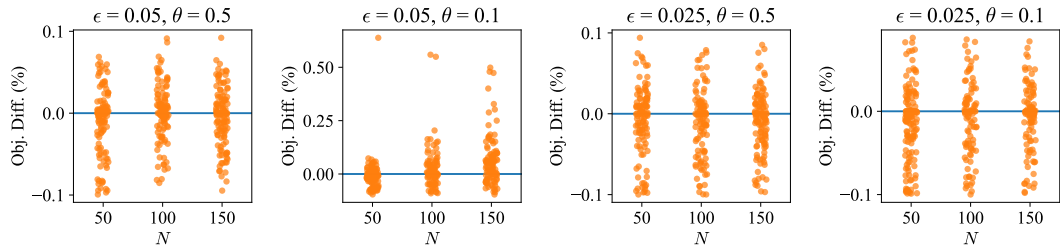


Figure 3.4: Performance comparison of the optimality for the proposed SFLA and the benchmark ExactS for the UC problem.

Each round dot represents a result of one of the 150 random runs. The blue horizontal line indicates a zero difference of the optimal value of a benchmark method compared to the proposed SFLA. Values higher than this horizontal line indicate that the benchmark achieves lower cost (better optimality) than the proposed SFLA. The optimisation horizon is set to $T = 24$.

Results for unit commitment

Also, the numerical results in this chapter suggest that both the proposed SFLA and ExactS exhibit variability. To reach a more robust conclusion, this section tests a range of parameter settings and implement 150 random runs for each setting. Figure 3.3 visualises the distribution of the *TimeF* metric, while the mean values and the computational speedup of the proposed SFLA over ExactS are summarised in Table 3.1. As can be seen, except for a single parameter setting ($\epsilon = 0.025, \theta = 0.1, N = 50$) where the proposed SFLA shows a minor underperformance (12% slower), the SFLA demonstrates significant computational speedup in *TimeF* over ExactS: the computational speedup reaches 10 $\times$ for a parameter setting ($\epsilon = 0.025, \theta = 0.1, N = 150$), and the speedup is 3.8 $\times$ for the average *TimeF* of the 12 parameter settings. Moreover, as shown in Figure 3.3, the upper bounds of the error bars (97.5-th percentile) for the proposed SFLA are lower than those of ExactS in 8 out of 12 parameter settings and comparable in the remaining settings, indicating a reduced likelihood of extreme cases with prolonged computation times.

ϵ	θ	N	<i>TimeF</i> (s)		<i>Time</i> (s)		<i>Reli.</i> (%)		<i>Obj. Diff.</i> (%)
			SFLA	ExactS	SFLA	ExactS	SFLA	ExactS	
0.05	0.5	50	118.17	643.19 (5.44×)	376.25	812.62 (2.16×)	99.14	99.20	−0.010
		100	88.04	343.21 (3.90×)	512.4	473.42 (0.92×)	99.67	99.68	0.001
		150	73.82	208.25 (2.82×)	471.25	288.92 (0.61×)	99.80	99.80	−0.004
	0.1	50	135.24	776.37 (5.74×)	576.6	1098.01 (1.90×)	76.01	76.88	−0.008
		100	91.07	437.21 (4.80×)	521.25	1064.36 (2.04×)	88.03	88.51	0.027
		150	185.28	273.93 (1.48×)	451.55	834.81 (1.85×)	91.87	92.14	0.061
0.025	0.5	50	326.1	525.47 (1.61×)	353.36	529.32 (1.50×)	100.00	100.00	−0.008
		100	148.47	754.92 (5.08×)	459.6	827.44 (1.80×)	100.00	100.00	−0.011
		150	111.05	449.19 (4.04×)	331.49	636.40 (1.92×)	100.00	100.00	−0.010
	0.1	50	159.67	141.29 (0.88×)	186.68	173.72 (0.93×)	87.81	87.75	−0.015
		100	120.77	949.39 (7.86×)	531.27	1079.66 (2.03×)	94.44	94.78	−0.009
		150	62.11	674.81 (10.86×)	605.29	1100.77 (1.82×)	96.45	96.71	−0.008

Table 3.1: Performance comparison of the final solution for the UC problem.

Each entry represents the mean value over 150 random runs. In columns *TimeF* (s) and *Time* (s), values in parenthesis represent the speed-up rate, i.e., the ratio of the time metrics of ExactS relative to the proposed SFLA.

Regarding the time required to achieve optimality (*Time*), Table 3.1 shows that the proposed SFLA demonstrates slower performance in 3 of 12 settings, indicating challenges in achieving the final minor solution improvements needed to meet the desired MIPGap. However, the speedup rate is still around 1.7× for the average *Time* of the 12 parameter settings, suggesting an overall advantage of the proposed SFLA over ExactS.

Figure 3.4 compares the quality of the optimal solution of the proposed SFLA with that of ExactS. In most random runs and most parameter settings, the *Obj. Diff.* is within the preset MIPGap, demonstrating comparable optimal solution quality between the proposed SFLA and ExactS. Slight underperformance (less than 0.6%) for certain random runs appears only under the combination of a high risk ($\epsilon = 0.05$), a small radius ($\theta = 0.1$), and a large number of samples ($N \geq 100$), which is consistent with Corollary 5 and the related discussions. However, among these underperformed parameter settings, the average underperformance is still below 0.061% as presented in Table 3.1. In addition, real-world applications such as those found in power systems have a small tolerance for risk, making the proposed SFLA an effective approximation method.

Appendix A.3.2 of this thesis further compares the proposed SFLA and ExactS in different problem scales (different T). It is observed that the proposed SFLA preserves computational efficiency in more complex cases without sacrificing optimality, making it particularly advantageous for practical high-complexity applications such as UC.

3.3.2 Bilevel storage strategic bidding problem

This section explores a bilevel optimisation problem that allows a price-maker energy storage operator (a typical DER) to maximise its profit by participating in a joint day-ahead energy and reserve market. The market participation process can be summarised as follows: Participants submit their bids and offers for each energy and reserve product for specific future delivery times. Bids represent the provision of services to the power system, such as generating power or supplying reserve capacity, while offers involve procuring services, such as purchasing energy for storage charging. In line with most established power markets, this section models each bid or offer as a price-quantity pair that is submitted to the market for clearing, which then determines a market clearing price (marginal price) and cleared quantities bounded by the submitted bid-offer quantities. Upon settlement, market participants receive payment or incur charges at the market clearing price for their respective cleared quantities, with a binding obligation to fulfill these quantities.

An energy storage operator transitions from a price-taker to a price-maker when its storage power capacity is large enough to influence market outcomes. To maximise profit, the storage operator determines the optimal bidding or offering strategy by solving a bilevel strategic bidding problem, where the lower level models the market clearing process [149, 151, 298, 299]. With increasing uncertainty due to renewable generation, the market clearing process can incorporate chance constraints (RHS-WDRJCC), such as those in the UC problem (see Section 3.3.1). Therefore, to accurately model the market clearing process, the bilevel problem needs to incorporate a RHS-WDRJCC in the lower level. Consequently, convex approximations to RHS-WDRJCC become important for transforming the bilevel problem into a solvable single-level counterpart via KKT conditions or strong duality conditions. Note that chance constraints at lower levels frequently arise in bilevel and multilevel optimisation problems, where uncertainty is revealed after lower-level decisions [300]. Another example is the European multilevel gas market clearing problem studied by [301].

This section compares the proposed SFLA with typical convex approximation benchmarks including LA [122] in Eq. (3.16), W-CVaR [111, 122] in Eq. (3.13), and Bonferroni approximation [122] in Eq. (3.10). For all the RHS-WDRJCC approximations, the corresponding bilevel problems are reformulated as single-level mixed-integer nonlinear non-convex programming problems, where the lower-level problem is replaced by the corresponding KKT optimality conditions, which can be solved using the Gurobi solver. ExactS is not included because bilevel problem solving requires convexity at the lower level. The bilevel formulation and also the corresponding single-level reformulation using KKT can be found in the online GitHub repository.

Evaluation metrics

The exclusive evaluation metrics used in this case study include:

1. *N Infeasible*: This is the number of random runs that are infeasible. This section introduces this metric because the high complexity of the bilevel problem introduces numerical issues that may make solvers incorrectly assess a feasible instance as infeasible. In the case studies of this chapter, this issue happens in certain runs where LA and W-CVaR are incorrectly assessed as infeasible, but the feasibility can be recovered by setting `NumericalFocus=3`, which increases numerical stability at the cost of longer solving time. A large value of this metric indicates that the method is prone to numerical issues, especially for LA and W-CVaR.
2. *Bilevel Profit* (k\$): This represents the optimal profit of storage systems, which is the objective obtained by solving the bilevel problem.
3. *Actual Profit* (k\$): The market clearing process is modeled as a single-level problem, where the market operator uses the exact MIP reformulation of RHS-WDRJCC to maximise social welfare, consistent with the ExactS approach in Section 3.3.1. Therefore, the use of ExactS may cause *Bilevel Profit* to differ from *Actual Profit*. The *Actual Profit* is calculated through the following steps: 1) Retrieve the bids and offers from the bilevel problem solutions; 2) Adjust bid prices slightly downward and offer prices slightly upward (by 10^{-5}), setting bid quantities to zero when the bilevel-cleared quantities are zero—these adjustments ensure the desired acceptance or rejection behaviors modeled in the bilevel problem; 3) Input these modified bids and offers into the actual market clearing problem with ExactS (MIP); 4) Calculate the actual market clearing price, which is then used to compute the *Actual Profit* of the storage operator. Due to the MIP formulation in ExactS, optimal dual variables cannot be directly used to determine marginal prices. Instead, sensitivity analysis is used to calculate the clearing price, by comparing the change of the objective function of the actual market clearing after a 10^{-3} MW increase in demand for each time step.
4. *Actual Profit Diff* (k\$): The difference of *Actual Profit* between a benchmark method and that calculated by the proposed SFLA. A negative value represents a lower profit compared to SFLA. This section does not use the percentage difference as is in the UC case because the profit could be close to zero, leading to exacerbated results. When calculating *Actual Profit Diff*, this section considers only those random runs where both the proposed SFLA and the benchmark identify feasible solutions within the prescribed `TimeLimit`.

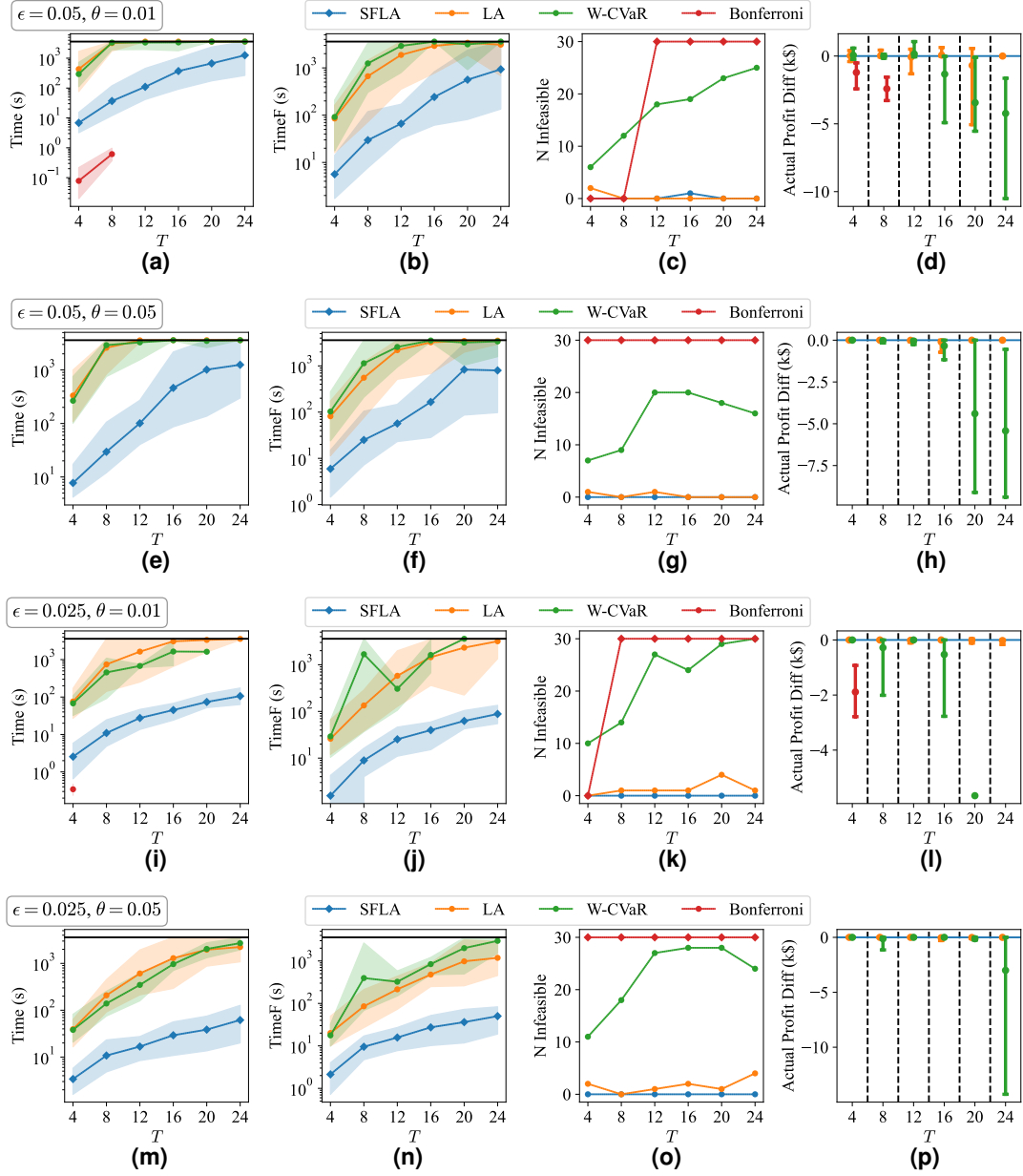


Figure 3.5: Performance comparison for the bilevel problem.

The evaluations are based on different parameter combinations of risk level (ϵ), radius (θ) parameters, and the number of time steps T . Subplots (e)–(p) replicate (a)–(d) but for varying (ϵ , θ) combinations. The black horizontal line represents the 3600s TimeLimit. Dots represent the mean values of the 30 random runs, with shaded areas representing the 95% percentile interval (from 2.5th to 97.5th). The blue horizontal line indicates a zero difference of the optimal value compared to the proposed SFLA. The *TimeF* (s) of Bonferroni approximation is not displayed because Bonferroni approximation is overly conservative such that it can easily find high *Bilevel Profit* (k\$) (the bilevel objective value) but poor *Actual Profit* (k\$) (plots 3.5d and 3.5i).

Results for bilevel strategic bidding

Figure 3.5 visualises the distributions of the evaluation metrics for the proposed SFLA and three benchmarks, i.e., LA, W-CVaR, and Bonferroni. As can be observed from the first two columns of Figure 3.5, the proposed SFLA demonstrates significant speedup in both *Time* and *TimeF*; to facilitate interpretation, Table A.1 in Appendix A.4 lists the mean values of these time metrics along with the corresponding speedup rates. In particular, the proposed SFLA achieves a speedup of up to 90× in solving to optimality (*Time*) compared to LA and W-CVaR, and the speedup is up to 40× in finding the first comparable high-quality solution (*TimeF*) relative to LA and 180× relative to W-CVaR. Although the Bonferroni approximation achieves the fastest computation, its over-conservativeness leads to infeasibility in the majority of parameter settings. It should be noted that the speedup rates compared to LA and W-CVaR decrease for a large T especially for $\epsilon = 0.05$; this is because LA and W-CVaR reach the *TimeLimit* rather than due to a reduction in the computational superiority of the proposed SFLA.

Furthermore, as can be observed in the third column of Figure 3.5, W-CVaR shows a large number of infeasible runs. These infeasibilities are resolved by setting `NumericalFocus=3`, indicating numerical issues. While LA is more numerically stable, it still incurs 22 infeasible runs in total. In contrast, the proposed SFLA has only one infeasible run, demonstrating its higher numerical stability. Note that the infeasibility of the Bonferroni approximation is caused by its over-conservativeness that cannot be recovered by setting `NumericalFocus=3`.

The fourth column of Figure 3.5 compares the *Actual Profit* of the benchmarks with that of the proposed SFLA. Among the feasible cases, Bonferroni approximation leads to significantly lower profit than other methods because of its over-conservativeness. It is noteworthy that both LA and W-CVaR can occasionally yield either higher or lower profits than the proposed SFLA. This variability arises because the market clearing modeled in the bilevel problem does not precisely replicate the actual market clearing process (cleared with exact MIP-based RHS-WDRJCC), and bilevel problems admit multiple solutions even at optimality. Thus, even when different methods yield the same *Bilevel Profit*, discrepancies can still arise in *Actual Profit*. It should be noted that W-CVaR shows significantly lower *Actual Profit* especially for a large T , because W-CVaR fails to find a good feasible solution within the *TimeLimit*, rather than due to over-conservativeness.

3.4 Concluding remarks

This chapter proposed a novel Strengthened and Faster Linear Approximation (SFLA) for RHS-WDRJCC. The proposed SFLA enables more efficient and less conservative solutions to a class of real-world large-scale chance-constrained optimisation problems, such as chance-constrained UC and bilevel optimisation for strategic bidding in electricity markets with chance-constrained ED. By applying valid inequalities to an existing linear approximation (LA) that is equivalent to the worst-case conditional value-at-risk (W-CVaR) method under specific hyperparameters, the proposed SFLA reduces the number of constraints and tightens the feasible region for the ancillary variables, resulting in a significant computational speedup. Despite this tightening, this chapter proves that the proposed SFLA does not introduce additional conservativeness and can even be a less conservative inner-approximation compared to LA. Furthermore, this chapter provides guidance in setting the hyperparameters of the proposed SFLA, specifically that the least conservativeness is achieved when hyperparameters are restricted to be uniform. This chapter also analyses conditions under which the proposed SFLA is equivalent to LA and W-CVaR.

Extensive numerical experiments demonstrate the superiority and significance of the proposed SFLA. In the UC problem, this chapter demonstrates that the proposed SFLA achieves up to $10\times$ and on average $3.8\times$ computational speedup compared to the existing exact and strengthened MIP reformulation (ExactS) in finding comparable high-quality solutions. This chapter also demonstrates that the proposed SFLA achieves approximation quality comparable to the exact reformulation ExactS, except for minor inferiority only under the combination of a high risk ϵ , a large number of historical data (N), and a small ambiguity set radius θ . In the bilevel strategic bidding problem where the exact formulation is not applicable due to its non-convexity, the proposed SFLA can lead to $90\times$ computational speedup compared to existing convex approximation methods (LA and W-CVaR). Note that, although the computational speed-up of SFLA is demonstrated on only two applications (UC and bilevel strategic bidding), each application is evaluated under multiple parameter settings, with 30–150 random runs using different generator parameters for varying risk levels and ambiguity set radii, as detailed in Appendices A.3.1 and A.4.1. These extensive evaluations provide a robust demonstration of the computational speedup of the proposed SFLA.

It is worth noting that the applications of the proposed SFLA are not limited to the UC and bilevel energy bidding problems demonstrated in this chapter. The method is also applicable to other large-scale, real-world problems that are already complex in their deterministic form and require tractable approaches to handle uncertainty. Examples include large-scale transportation problems, multilevel market clearing problems, and multilevel infrastructure planning, where the lower level makes decisions before uncertainty realisation. Finally, historical data plays a significant role that affects the complexity and solution quality of WDRJCC, including the proposed SFLA.

Recall that the first research question of this thesis is *Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering the DER uncertainty?* This Chapter addresses part of the first research question by tackling the complexity of general power system scheduling problems with RHS-WDRJCC. Although the proposed method has achieved significant computational speedup, the large number of DERs renders the scheduling problem intractable even in the deterministic case, as discussed in Sections 1.1.1, 2.1.1, and 2.1.2. A more efficient operational regime, such as DER aggregation, is required to make optimal DER flexibility utilisation computationally feasible. However, the introduction of such a new operational regime does not eliminate the challenges posed by DER uncertainty. It remains critical to integrate uncertainty management techniques, such as JCC, into the DER aggregation process, which is the focus of the next chapter.

Aggregating Distributed Energy Resources under Uncertainty

In line with the previous chapter, this chapter also focuses on the first research question of this thesis:

Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering the DER uncertainty?

The previous Chapter 3 addressed part of the first research question by tackling the complexity of general power system scheduling problems with RHS-WDRJCC. However, as discussed in Sections 1.1.1, 2.1.1, and 2.1.2, the large number of DERs renders the scheduling problem intractable even in deterministic settings, which necessitates alternative scheduling regimes, such as DER aggregation.

The key output of DER aggregation is the Aggregated Feasible Power Region (AFPR) as outlined in Section 2.1.2. The AFPR characterises the range of aggregated active and/or reactive power that DERs can provide while respecting network constraints [24, 25]. This region can be represented as a set of optimisation constraints and can be integrated into upper-level scheduling problems, which are more computationally tractable than direct scheduling [29]. This chapter specifically focuses on the aggregation of active power only, a subset of the AFPR referred to as the Aggregated Feasible Active Power Region (AFAPR). Notably, the proposed method in this chapter, though centred on active power aggregation, can be naturally extended to the more general AFPR.

As discussed in Sections 2.1.2 and 2.4.1, while DER aggregation is a promising approach, deriving the AFPR is NP-hard. Consequently, existing research has predominantly focused on approximation methods for the AFPR. However, these studies have largely relied on ICC in the derivation process, neglecting JCC. This results in unreliability (as will be demonstrated in the case study of this chapter) and highlights a significant research gap.

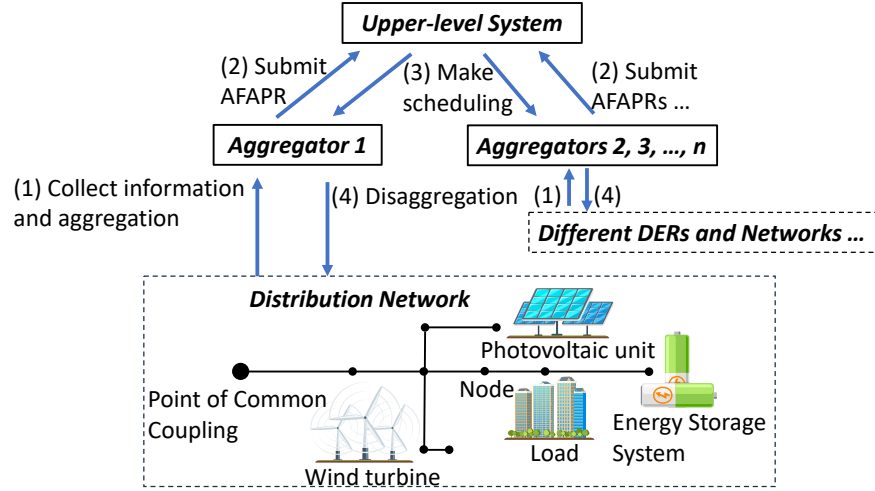


Figure 4.1: The scheduling process using the AFAPR.

This chapter addresses this gap by proposing a novel and computationally efficient method to integrate JCC in a two-stage setting into DER aggregation. The core idea is that, when uncertainty is not considered, the AFAPR can be approximated by a surrogate polytope [95, 102], which consists of a small set of linear constraints on aggregated power. This approximation implicitly accounts for DER constraints and network constraints. Starting from this surrogate polytope, a WDRJCC or JCC formulation becomes more tractable.

The proposed method supports both WDRJCC and classical JCC. For WDRJCC, the AFAPR can be exactly reformulated as a set of MILP constraints. For classical JCC, while there is no analytical expression for the AFAPR, the jointly reliable aggregated power can be determined through an iterative solution method. Importantly, both the MILP reformulation and the iterative process exhibit manageable computational complexity, independent of the number of DERs or the size of the network. Numerous case studies have demonstrated that the proposed method leads to a less conservative but still reliable AFAPR.

This chapter is organised as follows. Section 4.1 introduces the problem definition, and Section 4.2 outlines the core structure of the proposed method. Section 4.3 details the integration of WDRJCC into the AFAPR, reformulating it into tractable MILP constraints, while Section 4.4 presents the corresponding case study. Section 4.5 describes the method for integrating classical JCC to obtain the jointly reliable aggregated power using an iterative approach, with the corresponding case study provided in Section 4.6.

4.1 Problem definition

The process of DER scheduling via aggregation is depicted in Fig. 4.1. An aggregator manages a set of DERs including photovoltaic units (PV), wind turbines (WT), energy storage (ES), and controllable loads (CL) in a distribution network, which connects to the upper-level system through the Point of Common Coupling (PCC). The aggregator collects information about these DERs, and this thesis assumes that the aggregator has access to distribution network information (network model, line flow limits, and voltage limits), authorised by the distribution system operator (DSO) under some contracted conditions [302]. This access is important in the aggregation process to ensure reliable distribution network operation, and this importance can facilitate the collaboration between aggregators and the DSO. Also, the aggregator role may be done by the DSO itself to facilitate the coordination with the transmission network operator (TSO), known as the TSO-DSO coordination [303]. When the aggregator and the DSO are two separate entities, there could be conflicting objectives. For example, the aggregator seeks maximum profit, while the DSO aims at minimising system cost. In this case, additional coordination strategies must be in place to ensure fair consideration of different objectives [304, 305]. This chapter focuses primarily on the feasibility characterisation at the PCC that aggregates downstream DERs and the network, i.e., the AFAPR, which is a flexible and scalable concept that can also be used for aggregation at different nodes or for different parts of a distribution network [24], improving the computing efficiency of the coordination strategies.

There may be more than one aggregator located in the same downstream network. For each aggregator, the impact on network constraints by the power of invisible DERs owned by other aggregators can be modelled as uncertain non-controllable loads, as will be described in Section 4.1.1. It is also possible that the DSO as mediator collects all the information from the downstream aggregators and derives the reliable AFAPR based on the proposed method to facilitate upper-level scheduling and avoid the conservativeness to counter the additional uncertainty.

Given the DER and the network information, the AFAPR is derived and submitted to the upper-level system, which at the same time receives the AFAPRs from other aggregators for different DERs and networks, and performs a system-wide scheduling on these AFAPRs. The use of AFAPRs reduces computing efforts compared to directly scheduling all DERs. The private DER information and network details are also preserved by each aggregator. As the next step, each aggregator is informed of the scheduled aggregated power, which is disaggregated to individual DERs in the real-time implementation process. This chapter focuses on the derivation of the reliable AFAPR for the aggregated active power; while the reactive power aggregation, upper-level scheduling, and disaggregation are not the main scope. The proposed method can be applied to the AFPR that aggregates both active and reactive power, as will be discussed in Remark 3 in Section 4.1.3.

4.1.1 Model of distributed energy resources

Photovoltaic units and wind turbines

The PV operational power is modelled as

$$\forall t \in [T], \forall i_{PV} \in [N^{PV}],$$

$$0 \leq p_{t,i_{PV}}^{PV,\psi} \leq \bar{p}_{t,i_{PV}}^{PV,\psi}, \quad (4.1)$$

$$(p_{t,i_{PV}}^{PV,\psi}, q_{t,i_{PV}}^{PV,\psi}) \in \mathcal{PQ}_{t,i_{PV}}^{PV,\psi}. \quad (4.2)$$

WT has the same form of constraints which are not displayed. $\bar{p}_{t,i_{PV}}^{PV,\psi}$ is the maximum active power generation of the PV connected to bus i_{PV} at time step t , which depends on the weather and the capacity limit and is uncertain. Here, $\psi \in \Psi$ represents the phase index, where Ψ comprises a, b, c for Wye connections or ab, bc, ca for Delta connections. $[N^{PV}]$ collects the buses connected with PVs. The actual PV generation $p_{t,i_{PV}}^{PV,\psi}$ can be scheduled to zero as Eq. (4.1). PV and WT can also provide reactive power support $q_{t,i_{PV}}^{PV,\psi}$ and $q_{t,i_{PV}}^{WT,\psi}$. Their reactive power constraints are described by their PQ-charts [10] in Eq. (4.2), which is illustrated as the area surrounded by the capacity limit (green curves), active power limits (red curves), and reactive power limits (blue curves) in Fig. 4.2. To address nonconvexity in PQ charts, this thesis converts them into sets of linear constraints (grey areas) following [94].

Energy storage

The operational power of an ES is modelled as:

$$\forall t \in [T], \forall i_{ES} \in [N^{ES}],$$

$$p_{t,i_{ES}}^{ES,\psi} = \hat{p}_{t,i_{ES}}^{ES,\psi} + \check{p}_{t,i_{ES}}^{ES,\psi}, \quad (4.3)$$

$$\underline{p}_{i_{ES}}^{ES,\psi} \leq \check{p}_{t,i_{ES}}^{ES,\psi} \leq 0, \quad (4.4)$$

$$0 \leq \hat{p}_{t,i_{ES}}^{ES,\psi} \leq \bar{p}_{i_{ES}}^{ES,\psi}, \quad (4.5)$$

$$e_{t,i_{ES}} = e_{t-1,i_{ES}} + \sum_{\psi} (\hat{p}_{t,i_{ES}}^{ES,\psi} \hat{\eta} + \check{p}_{t,i_{ES}}^{ES,\psi} / \check{\eta}) \cdot \Delta t, \quad (4.6)$$

$$\underline{e}_{i_{ES}} \leq e_{t,i_{ES}} \leq \bar{e}_{i_{ES}}, \quad (4.7)$$

$$(p_{t,i_{ES}}^{ES,\psi}, q_{t,i_{ES}}^{ES,\psi}) \in \mathcal{PQ}_{t,i_{ES}}^{ES,\psi}. \quad (4.8)$$

Here, $[N^{ES}]$ collects the buses connected with ESs, and $p_{t,i_{ES}}^{ES,\psi}$ is the power of the ES connected to bus i_{ES} at phase ψ and time step t . The ES has separately modelled charging power $\hat{p}_{t,i_{ES}}^{ES,\psi}$ and discharging power $\check{p}_{t,i_{ES}}^{ES,\psi}$, subject to the charging power limit $\bar{p}_{i_{ES}}^{ES,\psi}$ and discharging limit $\underline{p}_{i_{ES}}^{ES,\psi}$ as Eqs. (4.4)-(4.5). The charging and discharging affect the energy level $e_{t,i_{ES}}$ as the time-coupling constraint (4.6), which is subject to the charging (discharging) efficiency $\hat{\eta}$ ($\check{\eta}$) due to the power loss. Δt is the length of one time step. The energy level of an ES is limited within the maximum energy $\bar{e}_{i_{ES}}$ and the minimum energy $\underline{e}_{i_{ES}}$ (4.7). The ES's PQ chart is

also linearised in Fig. 4.2. Note that simultaneous charging and discharging can be avoided by adding a linear penalty term $\omega_{iES} \sum_{t,\psi,iES} (\hat{p}_{t,iES}^{ES,\psi} - \check{p}_{t,iES}^{ES,\psi})$ to the scheduling objective with ω_{iES} being a preset penalty coefficient [100]. This penalty also represents the linear cost of battery degradation with respect to the throughput. This thesis models this cost in both the aggregated cost function and the disaggregation process in Section 4.3.2. Ref. [306] proved this result for $\forall \omega_{iES} > 0$ in a DCOPF problem under some common power system statuses via KKT conditions.

Load

This thesis considers CLs that can adjust their power within a continuous range modelled as

$$\forall t \in [T], \forall i_{CL} \in [N^{CL}],$$

$$\sigma_{\min} \ddot{p}_{t,i_{CL}}^{CL,\psi} \leq p_{t,i_{CL}}^{CL,\psi} \leq \sigma_{\max} \ddot{p}_{t,i_{CL}}^{CL,\psi}, \quad (4.9)$$

$$\sum_t^T p_{t,i_{CL}}^{CL,\psi} = \sum_t^T \ddot{p}_{t,i_{CL}}^{CL,\psi}, \quad (4.10)$$

$$q_{t,i_{CL}}^{CL,\psi} = \alpha^{CL} \cdot p_{t,i_{CL}}^{CL,\psi}, \quad (4.11)$$

where $\ddot{p}_{t,i_{CL}}^{CL,\psi}$ is the uncertain base load of the CL without any intervention. Parameters $\sigma_{\min}$ and $\sigma_{\max}$ model the minimum and maximum power levels (relative to the base load) that the CL can be controlled to. Eq. (4.10) ensures that the total consumed energy equals the original energy consumption. This thesis assumes that the CL has a fixed power factor α^{CL} as Eq. (4.11). There are also uncertain non-controllable loads (NL) in the network with $\sigma_{\min} = \sigma_{\max} = 1$. Note that this chapter focuses on the case where a single aggregator collects all the DER information and the network information of a distribution network. However, it is possible to extend to the multiple-aggregator-single-network case, where other aggregators' power scheduling can be regarded as part of the uncertain NLs modelled here.

Note that the DER models in this chapter represent simplified modeling. For example, the ES has charging and discharging efficiency that nonlinearly depends on the charging and discharging power [307]. There is battery degradation, which is also a nonlinear process with respect to time and cycle numbers [308]. There are also extra considerations for CL, such as the uncertain users' responsiveness and response accuracy. However, incorporating all these accurate models leads to complicated aggregation problems that are challenging to handle, as discussed in Section 2.1.2. Deriving more accurate yet efficient DER models, or designing more computationally efficient aggregation methods for the accurate models, is an important future research direction.

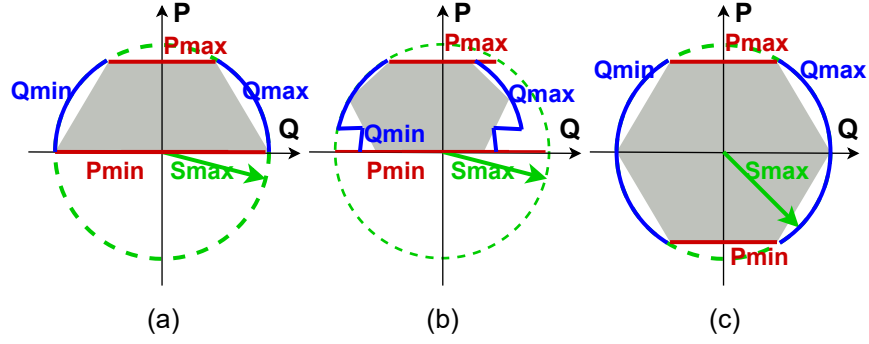


Figure 4.2: The PQ capability chart for (a) PV, (b) WT, and (c) ES.

These charts are based on [10].

4.1.2 Network model

Following Refs. [309] and [310], this thesis considers a generic three-phase distribution network with one slack bus (the PCC in Fig. 4.1) and N^B three-phase PQ buses. Denote by s_t^{Agg} the three-phase complex power injection of the slack bus (PCC), by s_t^Y the three-phase complex power injection of all the PQ buses from Wye sources, by s_t^D the three-phase complex power injection of all the PQ buses from Delta sources, by v_t^0 the three-phase complex nodal voltage of the slack bus, by v_t the three-phase complex nodal voltages of all the PQ buses, by i_t the phase net current injections of all the PQ buses, and by i_t^D the phase-to-phase currents of all the PQ buses. Furthermore, denote the complex three-phase bus admittance matrix as:

$$Y := \begin{bmatrix} Y_{00} & Y_{0L} \\ Y_{L0} & Y_{LL} \end{bmatrix} \in \mathbb{C}^{3(N^B+1) \times 3(N^B+1)}, \quad (4.12)$$

where $\mathbb{C}$ is the complex domain. Y_{00} , Y_{0L} , Y_{L0} , and Y_{LL} indicate self-/mutual- admittance matrices of/between slack/non-slack buses. The bus admittance matrix Y can be formed from the network topology, the π -model of the three-phase power lines, and other passive elements, representing the physical coupling of a three-phase distribution network. Based on [309], the ACPF equations can be expressed as:

$$\text{diag}(\mathbf{H}^\top \overline{i_t^D}) v_t + s_t^Y = \text{diag}(v_t) \overline{i_t}, \quad (4.13)$$

$$s_t^D = \text{diag}(\mathbf{H} v_t) \overline{i_t^D}, \quad (4.14)$$

$$i_t = Y_{L0} v_t^0 + Y_{LL} v_t, \quad (4.15)$$

$$s_t^{\text{Agg}} = \text{diag}(v_t^0) (\overline{Y_{00} v_t^0} + \overline{Y_{0L} v_t}), \quad (4.16)$$

where $\overline{i_t^D}$ indicates conjugate, and $\text{diag}(\cdot)$ constructs a square matrix with the elements of the vector inside as its diagonal. $\mathbf{H}$ is a coefficient matrix. Refer to [309, 310] for a detailed interpretation.

The ACPF-based network constraints are nonlinear and nonconvex, which hinders the calculation of AFPR. To tackle this challenge, this thesis applies the linear distribution network model in [309]. By interpreting the fixed-point function derived from the AC power-flow equations, [309] showed that the three-phase voltage magnitudes $|v_t|$, the three-phase line current magnitudes $|i_t^L|$, and the aggregated active power at the PCC p_t^{Agg} (summing over the three phases) can be expressed by linear functions w.r.t. the three-phase active (the real parts of s_t^Y and s_t^D) and reactive (the imaginary parts of s_t^Y and s_t^D) power injections of the N^B PQ buses. Let x_t^{DER} collect the active and reactive power of DERs (decision variables) plus NLs (uncertain parameters) at all the connected phases and buses at time step t :

$$x_t^{\text{DER}} := [p_{t,i_d}^{d,\psi}, q_{t,i_d}^{d,\psi}, \ddot{p}_{t,i_{\text{NL}}}^{\text{NL},\psi}, \alpha^{\text{NL}} \ddot{p}_{t,i_{\text{NL}}}^{\text{NL},\psi}] \quad (4.17)$$

$$i_d \in [N^d], d \in \{\text{PV}, \text{WT}, \text{ES}, \text{CL}\}, i_{\text{NL}} \in [N^{\text{NL}}], \psi \in \Psi.$$

By combining the linear relationship between x_t^{DER} and the real and imaginary parts of s_t^Y and s_t^D , this thesis has the following linear network model:

$$\forall t \in [T], |i_t^L| = J_t x_t^{\text{DER}} + j_t^0, \quad (4.18)$$

$$|v_t| = K_t x_t^{\text{DER}} + k_t^0, \quad (4.19)$$

$$p_t^{\text{Agg}} = g_t^\top x_t^{\text{DER}} + g_t^0, \quad (4.20)$$

$$|i_{\min}^L| \leq |i_t^L| \leq |i_{\max}^L|, \quad (4.21)$$

$$|v_{\min}| \leq |v_t| \leq |v_{\max}|. \quad (4.22)$$

where $J_t, j_t^0, K_t, k_t^0, g_t, g_t^0$ are linear coefficients derived from a given operational point, a no-load case, and the three-phase bus admittance matrix Y , leading to better global performance than the first-order Taylor's approximation [309]. Refer to [309] for a detailed mathematical deduction. Its global performance is also theoretically analysed and verified through extensive case studies in [309, 310], showing a less than 0.6% (1.4%) voltage relative error for a 2000-node distribution feeder (IEEE-8500 feeder, resp.) under a range of load injections. Note that, in general, the accuracy of a single linear network model only holds when the optimised operating point is close to the given operating point used to derive the linear model coefficients [310, 311], given the nonlinear nature of the ACPF equations. Iterative approaches can be necessary to find accurate network solutions with updated linear model coefficients in each iteration [311].

Another promising track to deal with the computational challenge of the exact ACPF network model is convex relaxation. However, existing convex relaxation is only exact for specific conditions that are unfortunately difficult to predict in practice; when the conditions do not hold, there may be a large optimality gap and it is difficult to reconstruct a feasible solution [311–313]. In contrast, although there is no guarantee of exactness as well, the linear network model is much more computationally efficient given the current mature and reliable linear programming solvers [312].

The line current and nodal voltage need to be constrained within a given range as Eq. (4.21) and (4.22). Real-world distribution networks are typically voltage-constrained and the line current limits can be disregarded [28], especially for distribution networks with long power lines [314], so this thesis only considers the three-phase nodal voltage limits as Eq. (4.22). Incorporating the line current constraints is straightforward as it is also linear.

4.1.3 Problem target: jointly reliable AFAPR

The outcome of a DER aggregation problem is the AFAPR Ω^{Agg} , which is a feasible region for the aggregated active power $\mathbf{p}^{\text{Agg}} := [p_{t=0}^{\text{Agg}}, \dots, p_{t=T}^{\text{Agg}}]$. The aggregated power is coupled with the individual DER power and NL power $\mathbf{x}^{\text{DER}} := [\mathbf{x}_{t=0}^{\text{DER}}, \dots, \mathbf{x}_{t=T}^{\text{DER}}]$ by Eq. (4.20), which are further subject to the DER operational constraints and network constraints. Therefore, the AFAPR can be defined as:

$$\Omega^{\text{Agg}} = \{ \mathbf{p}^{\text{Agg}} \mid \exists \mathbf{x}^{\text{DER}}, \text{ Eqs. (4.1)–(4.22)} \forall \psi, i_d, t \}. \quad (4.23)$$

Fig. 4.3a visualises the proposed AFAPR. The feasible region defined by all the DER constraints and network constraints (4.1)–(4.22) for the $[\mathbf{p}^{\text{Agg}}, \mathbf{x}^{\text{DER}}]$ space can be represented as the high-dimension yellow polytope. Because the AFAPR Ω^{Agg} in (4.23) focuses on the aggregated power only, it can be represented as the red projected polytope in the $\mathbf{p}^{\text{Agg}}$ space in Fig. 4.3a. Note that (4.23) defines the AFAPR across T time steps rather than for one time step only, which ensures the time-coupling constraints of ESs and the energy requirements of CLs.

The uncertain DERs make the AFAPR Ω^{Agg} uncertain in the scheduling phase. Here, the uncertain parameters include the maximum active power capability of PV $\bar{p}_{t,i_{\text{PV}}}^{\text{PV},\psi}$ and WT $\bar{p}_{t,i_{\text{WT}}}^{\text{WT},\psi}$, the base load of CL $\bar{p}_{t,i_{\text{CL}}}^{\text{CL},\psi}$ and NL $\bar{p}_{t,i_{\text{NL}}}^{\text{NL},\psi}$, and thus a part of $\mathbf{x}^{\text{DER}}$. For managing the risk, it is desired to find a “jointly reliable AFAPR” $\Omega^{\text{Agg}}(\epsilon)$ such that any $\mathbf{p}^{\text{Agg}}$ within it can be implementable without violating any constraints for the entire scheduling horizon $[T]$ with the desired probability $1 - \epsilon$. This need can be tackled via a two-stage framework with a JCC. More specifically, the desired (maximum-volume) jointly reliable AFAPR $\Omega^{\text{Agg}}(\epsilon)$ can be

defined as:

$$\Omega^{\text{Agg}}(\epsilon) = \left\{ \mathbf{p}^{\text{Agg}} \mid \mathbb{P} \left(\mathbf{p}^{\text{Agg}} \in \Omega^{\text{Agg}} \right) \geq 1 - \epsilon \right\} \quad (4.24a)$$

$$= \left\{ \mathbf{p}^{\text{Agg}} \mid \mathbb{P} \left(\exists \mathbf{x}^{\text{DER}}, \text{ Eqs. (4.1)–(4.22)} \forall \psi, i_d, t \right) \geq 1 - \epsilon \right\}. \quad (4.24b)$$

The probability function $\mathbb{P}$ in Eq. (4.24b) defines a JCC with a two-stage structure, which ensures the $(1 - \epsilon)$ -probability existence of the second-stage DER disaggregation $\mathbf{x}^{\text{DER}}$ that meets all constraints in Ω^{Agg} simultaneously, which are the DER constraints and network constraints for all buses, phases, and future time steps.

The JCC in the jointly reliable AFAPR $\Omega^{\text{Agg}}(\epsilon)$ (4.24) is complicated as it involves a large number of DER constraints and network constraints and is under a two-stage setting, which necessitates approximations. For the safety preference in power system operations, this thesis specifically focuses on an inner-approximation of the desired $\Omega^{\text{Agg}}(\epsilon)$. In other words, the objective of this chapter is to find an approximated AFAPR $\Omega_{\text{Approx}}^{\text{Agg}}(\epsilon)$ such that:

$$\Omega_{\text{Approx}}^{\text{Agg}}(\epsilon) \subseteq \Omega^{\text{Agg}}(\epsilon) \text{ in (4.24)}. \quad (4.25)$$

The derived approximation $\Omega_{\text{Approx}}^{\text{Agg}}(\epsilon)$ also needs to have a mild complexity when it is integrated into the upper-level scheduling and be capable of most of the scheduling problems. The inner-approximation should not be overly conservative as well.

The AFAPR approximation problem can be alternatively defined as maximising the volume of the approximation, with (4.25) included in the constraints, which is similar to existing studies in [25, 160, 164]. However, the volume maximisation is hard to solve, because the targeted jointly reliable AFAPR $\Omega^{\text{Agg}}(\epsilon)$ in (4.24) is generally nonconvex and it is hard to find the analytic form due to the complicated JCC in a two-stage setting [105]. In addition, volume maximisation problems are only solvable for specific approximations such as boxes [25], ellipsoids [164], or shifting and scaling prototype polytopes, i.e., the homothet with the maximum volume [160]. Therefore, the output in the maximum-volume framework may not necessarily be less conservative due to the restriction on approximations.

Remark 1: It is important to ensure the reliability of AFAPR, i.e., high probability of implementability of the scheduled aggregated power $\mathbf{p}^{\text{Agg}}$, which is the main focus of this chapter. When a scheduled $\mathbf{p}^{\text{Agg}}$ is not implementable, there can be two implications: 1) DERs have enough power flexibility but distribution network constraints are violated, and/or 2) DERs do not have enough power flexibility. The former case threatens the network operation. The latter result may be compensated by the transmission grid, which, however, would become less viable when in the future the DSO plays a larger share in the whole system.

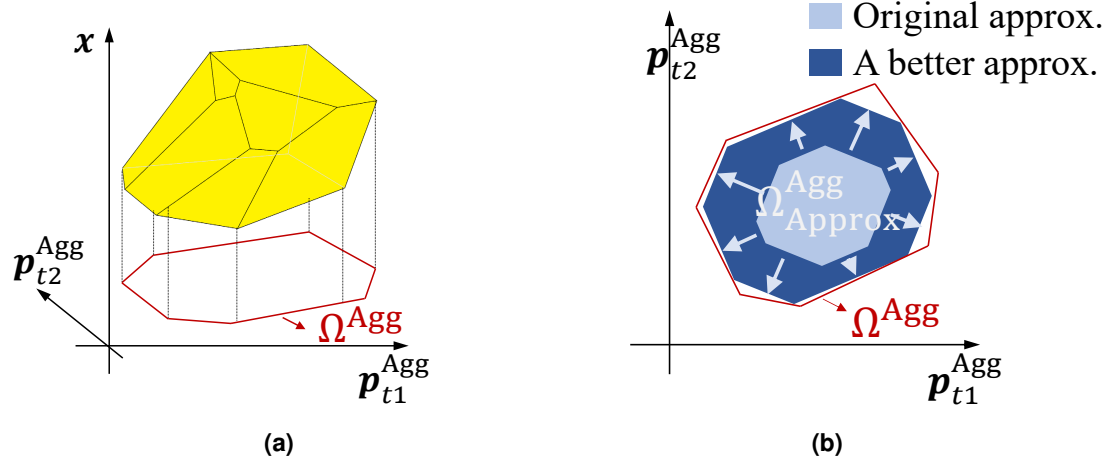


Figure 4.3: Illustration of AFAPR and its approximation.

(a): The high-dimension yellow polytope represents the feasible region of $[p^{\text{Agg}}, x^{\text{DER}}]$ defined by all the DER constraints and network constraints. Because the AFAPR Ω^{Agg} in (4.23) focuses on the aggregated power only, it can be represented as the red projected polytope in the p^{Agg} space. (b): A better inner-approximation can be achieved by tuning the intercept of $\Omega_{\text{Approx}}^{\text{Agg}}$ with the same slope matrix A^{PA} .

Remark 2: Depending on specific problems, an AFAPR containing unreliable power trajectories may still lead to reliable optimal solution to a scheduling problem. Therefore, one may explore the application-oriented/objective-dependent AFAPR based on information about the specific scheduling problems that will be carried out. Nonetheless, the focus of this chapter, namely $\Omega^{\text{Agg}}(\epsilon)$, is more universal as any scheduling problems based on it can lead to reliable solutions. This is particularly beneficial in TSO-DSO interactions where a DSO may not know the exact upper-level scheduling problem formulation before submitting the AFAPR.

Remark 3: Reactive power is also an important dimension in DER aggregation, that is, the AFPR or the PQ chart. Ref. [24] provided a systematic framework for the aggregated PQ chart at a single time step. Note that, since the time-coupling constraints of storage-like DERs couple the active power across time and the active power couples the reactive power, it may be necessary to have a time-coupled AFPR beyond the common single-time-step PQ aggregation [24]. The proposed method can also be applied to the time-coupled AFPR by slightly modifying the AFAPR definition in (4.23) and (4.24), where p^{Agg} will be replaced with $[p^{\text{Agg}}, q^{\text{Agg}}]$ and $q^{\text{Agg}} := [q_{t=0}^{\text{Agg}}, \dots, q_{t=T}^{\text{Agg}}]$.

4.2 Core framework of the proposed method

It is challenging to find a good inner-approximation under the complicated JCC in $\Omega^{\text{Agg}}(\epsilon)$ (4.24). However, starting from the surrogate polytope inner-approximation in a deterministic setting, this thesis instead applies the JCC to the surrogate polytope and finally derive an analytical expression for the inner-approximation of the jointly reliable AFAPR $\Omega^{\text{Agg}}(\epsilon)$ in a computationally tractable manner.

4.2.1 Surrogate polytope in a deterministic setting

When there is no uncertainty, it is noticed that the AFAPR Ω^{Agg} in Eq. (4.23) admits an inner-approximation $\Omega_{\text{Approx}}^{\text{Agg}}$ based on a “surrogate polytope” with a pre-selected shape [102], such that:

$$\Omega_{\text{Approx}}^{\text{Agg}} \subseteq \Omega^{\text{Agg}}, \quad \Omega_{\text{Approx}}^{\text{Agg}} = \{\mathbf{p}^{\text{Agg}} \mid \mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}} \leq \mathbf{b}^{\text{PA}}\}, \quad (4.26)$$

where the matrix $\mathbf{A}^{\text{PA}}$ controls the slope of each facet of the surrogate polytope approximation, i.e., shape, and is pre-selected, and the vector $\mathbf{b}^{\text{PA}}$ determines the intercept of each facet of the polytope approximation and is to be inferred. Fig. 4.3b shows that only adjusting the intercept $\mathbf{b}^{\text{PA}}$ while keeping the shape matrix $\mathbf{A}^{\text{PA}}$ unchanged can transfer the original approximation to a better approximation with a greater size (less conservative inner-approximation of the true AFAPR Ω^{Agg}). The parameters of $\mathbf{b}^{\text{PA}}$ for such a better polytope approximation can be inferred in a computationally affordable manner by methods such as in [25, 94, 95, 100, 102], where they also empirically showed the good approximation accuracy. As the network constraints and DER constraints are included in Ω^{Agg} , the polytope approximation also implicitly includes these low-level constraints due to $\Omega_{\text{Approx}}^{\text{Agg}} \subseteq \Omega^{\text{Agg}}$. The selection of $\mathbf{A}^{\text{PA}}$ is problem-dependent [94, 95, 102]. For example, if the DERs to be aggregated have energy constraints (such as energy or thermal storages), then $\mathbf{A}^{\text{PA}}$ can be selected to be the virtual battery shape, and its approximation quality has been empirically verified [95, 102].

Virtual battery approximation

This thesis specifically introduces the virtual battery (VB) approximation [95] as VB is used in the case studies. Note that the proposed method in this chapter works for arbitrary surrogate polytope approximations with pre-selected slopes. A VB model is essentially a set of power constraints and energy constraints similar to a battery:

$$\Omega_{\text{Approx}}^{\text{Agg}} = \{\mathbf{p}^{\text{Agg}} \mid \mathbf{A}^{\text{VB}} \mathbf{p}^{\text{Agg}} \leq \mathbf{b}^{\text{VB}}\}, \quad (4.27)$$

$$\mathbf{A}^{\text{VB}} = [\mathbf{I}, -\mathbf{I}, \mathbf{\Lambda}, -\mathbf{\Lambda}]^{\top}, \quad (4.28)$$

$$\mathbf{b}^{\text{VB}} = [\bar{p}_{t=1}^{\text{VB}}, \dots, \bar{p}_{t=T}^{\text{VB}}, -\underline{p}_{t=1}^{\text{VB}}, \dots, -\underline{p}_{t=T}^{\text{VB}}, \bar{e}_{t=1}^{\text{VB}}, \dots, \bar{e}_{t=T}^{\text{VB}}, -\underline{e}_{t=1}^{\text{VB}}, \dots, -\underline{e}_{t=T}^{\text{VB}}], \quad (4.29)$$

where $\mathbf{I} \in \mathbb{R}^{T \times T}$ is the identity matrix and $\mathbf{\Lambda} \in \mathbb{R}^{T \times T}$ is a lower-triangle matrix with all non-zero elements equal to 1. $\mathbf{b}^{\text{VB}}$ collects the VB parameters including the maximum power $\bar{p}_t^{\text{VB}}$, minimum power $\underline{p}_t^{\text{VB}}$, maximum energy $\bar{e}_t^{\text{VB}}$, and minimum energy levels $\underline{e}_t^{\text{VB}}$ at each time step $t \in [T]$.

The surrogate polytope constraints in (4.26) or the special case (4.27) apply to $\mathbf{p}^{\text{Agg}}$ only. Therefore, the number of constraints in (4.26) can be significantly smaller than that in Ω^{Agg} (4.23). Specifically, for the VB polytope, the number of constraints is only $4T$, leading to tractable integration of WDRJCC and JCC as will be introduced next.

4.2.2 Extension to jointly reliable AFAPR approximation

When considering the DER uncertainty, similar to the derivation of the jointly reliable AFAPR $\Omega^{\text{Agg}}(\epsilon)$ (4.24), a jointly reliable polytope approximation $\Omega_{\text{Approx}}^{\text{Agg}}(\epsilon)$ can be defined as:

$$\begin{aligned} \Omega_{\text{Approx}}^{\text{Agg}}(\epsilon) &= \{\mathbf{p}^{\text{Agg}} \mid \mathbb{P}(\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}) \geq 1 - \epsilon\} \\ &= \{\mathbf{p}^{\text{Agg}} \mid \mathbb{P}(\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}} \leq \tilde{\mathbf{b}}^{\text{PA}}) \geq 1 - \epsilon\}, \end{aligned} \quad (4.30)$$

where $\mathbb{P}(\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}} \leq \tilde{\mathbf{b}}^{\text{PA}}) \geq 1 - \epsilon$ is called “surrogate polytope JCC”. $\mathbb{P}$ is the probability function for the true distribution of $\tilde{\mathbf{b}}^{\text{PA}}$: since the slope $\mathbf{A}^{\text{PA}}$ is pre-selected, the uncertainty towards the approximated AFAPR (4.26) reduces to $\mathbf{b}^{\text{PA}}$ only, which can be modelled as a random vector $\tilde{\mathbf{b}}^{\text{PA}}$, highlighted by the overhead “tilde”. Also, given $\Omega_{\text{Approx}}^{\text{Agg}} \subseteq \Omega^{\text{Agg}}$ as defined in Eq. (4.26), the form of Eq. (4.30) satisfies the objective of this chapter in Eq. (4.25), i.e., an inner-approximation to the jointly reliable AFPR:

Proposition 2. $\Omega_{\text{Approx}}^{\text{Agg}}(\epsilon) \subseteq \Omega^{\text{Agg}}(\epsilon)$.

Proof. For any $\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}(\epsilon)$, the following can be derived:

$$\begin{aligned} &\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}(\epsilon) \\ \Leftrightarrow &\mathbb{P}(\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}) \geq 1 - \epsilon \\ \Rightarrow &\mathbb{P}(\mathbf{p}^{\text{Agg}} \in \Omega^{\text{Agg}}) \geq 1 - \epsilon \text{ due to } \Omega_{\text{Approx}}^{\text{Agg}} \subseteq \Omega^{\text{Agg}} \\ \Leftrightarrow &\mathbf{p}^{\text{Agg}} \in \Omega^{\text{Agg}}(\epsilon). \end{aligned} \quad (4.31)$$

Since the derivation above holds for any $\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}(\epsilon)$, it follows that $\Omega_{\text{Approx}}^{\text{Agg}}(\epsilon) \subseteq \Omega^{\text{Agg}}(\epsilon)$. $\square$

In this way, the two-stage structure of the JCC in $\Omega^{\text{Agg}}(\epsilon)$ (4.24) is implicitly considered through the “single-stage” surrogate polytope JCC in Eq. (4.30).

Furthermore, the proposed method leads to a special JCC structure in the jointly reliable polytope approximation Eq. (4.30), that is, JCC with RHS uncertainty. As reviewed in Section 2.2.1, this structure leads to tractable analytical reformulations or solution schemes. In the subsequent content, this chapter will demonstrate that the proposed method leads to both tractable integration of WDRJCC and classical JCC. As the surrogate polytope JCC is defined only for the aggregated power p^{Agg} , the computational complexity of the integration of WDRJCC and JCC will not increase with the number of DERs nor with the size of the network.

Remark 4: The key part of the proposed method is the derivation of the inner polytope approximation in a deterministic setting. In other words, as long as there are methods to calculate the polytope approximation in a deterministic setting, the proposed method can serve as an extension to integrate the complicated JCC with a two-stage structure to enhance reliability. Therefore, for other challenges such as non-convex network models and DER models, one can simply focus on the derivation in a deterministic setting, and then naturally use the proposed method in this chapter to further tackle uncertainty. Note that deriving the AFPR approximation involves a trade-off between computational complexity and approximation accuracy, as the exact calculation of AFPR is NP-hard, as discussed in Section 2.1.2. However, the key to note is that the proposed method for integrating JCC does not introduce any additional computational bottleneck.

4.3 Part I: integrating WDRJCC

This chapter first shows that the proposed method leads to tractable integration of WDRJCC.

The motivation of using WDRJCC is that the true distribution of $\tilde{\mathbf{b}}^{\text{PA}}$ is unknown and a subjective assumption can lead to limited out-of-sample performance, as has been extensively discussed in the previous chapter and Section 2.2.1. To represent the ambiguity towards the distribution of $\tilde{\mathbf{b}}^{\text{PA}}$, the polytope JCC in (4.30) is replaced with a distributionally robust form by WDRJCC; hence the WDRJCC-based approximation to $\Omega^{\text{Agg}}(\epsilon)$, denoted as $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$, can be defined as:

$$\begin{aligned} \Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta) &= \{ \mathbf{p}^{\text{Agg}} \mid \mathbb{P}'(\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}) \geq 1 - \epsilon, \forall \mathbb{P}' \in \mathcal{F}_N(\theta) \} \\ &= \{ \mathbf{p}^{\text{Agg}} \mid \mathbb{P}'(\mathbf{A} \mathbf{p}^{\text{Agg}} \leq \tilde{\mathbf{b}}^{\text{PA}}) \geq 1 - \epsilon, \forall \mathbb{P}' \in \mathcal{F}_N(\theta) \}, \end{aligned} \quad (4.32)$$

where $\mathcal{F}_N(\theta)$ is the Wasserstein ambiguity set as has been defined in Eq. (3.2) in Chapter 3. Other forms of ambiguity set exist. The Wasserstein set is selected for its superior out-of-sample guarantees [165].

4.3.1 Exact reformulation

This section then shows that the approximation $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ admits a tractable reformulation. Suppose N realisations have been collected for the random vector $\tilde{\mathbf{b}}^{\text{PA}} \in \mathbb{R}^P$ and denote the dataset for these realisations by $\mathcal{D} := \{\mathbf{b}^1, \dots, \mathbf{b}^N\}$, where P is the dimension of the random vector $\tilde{\mathbf{b}}^{\text{PA}}$ in Eq. (4.30), which also equals to the number of constraints within the WDRJCC in Eq. (4.32). Let q_p denote the $(\lfloor \epsilon N \rfloor + 1)$ -th smallest value of $\{b_{i,p}^{\text{PA}}, \dots, b_{N,p}^{\text{PA}}\}$, where $\lfloor \cdot \rfloor$ is the floor function and $b_{i,p}^{\text{PA}}$ denotes the p -th entry of the i -th realisation $\mathbf{b}_i^{\text{PA}}$ of $\tilde{\mathbf{b}}^{\text{PA}}$. Similar to Chapter 3, this chapter also introduces the index set $[N]_p$ as $[N]_p := \{i \in [N] : b_{i,p}^{\text{PA}} < q_p\}$ with $p \in [P]$.

Based on [108] (the ExactS method in Chapter 3), the WDRJCC in (4.32) admits an exact reformulation; thus the approximated reliable AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ can be equivalently written as a set of MILP constraints:

$$\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta) = \left\{ \mathbf{p}^{\text{Agg}} \in \mathcal{X}^{\text{Agg}} : \begin{cases} \exists \mathbf{z} \in \{0, 1\}^N, s \geq 0, \mathbf{r} \geq \mathbf{0}, & \text{(a)} \\ \epsilon N s - \sum_{i \in [N]} r_i \geq \theta N, & \text{(b)} \\ M(1 - z_i) \geq s - r_i, i \in [N], & \text{(c)} \\ \sum_{i \in [N]} z_i \leq \lfloor \epsilon N \rfloor, & \text{(d)} \\ b_{i,p}^{\text{PA}} - \mathbf{a}_p^\top \mathbf{p}^{\text{Agg}} + (q_p - b_{i,p}^{\text{PA}})z_i \geq s - r_i, \quad i \in [N]_p, p \in [P], & \text{(e)} \\ q_p - \mathbf{a}_p^\top \mathbf{p}^{\text{Agg}} \geq s, \quad p \in [P] & \text{(f)} \end{cases} \right\}, \quad (4.33)$$

where $\mathbf{z} \in \mathbb{B}^N$, $s \in \mathbb{R}$, $\mathbf{r} \in \mathbb{R}^N$ are auxiliary decision variables, and $\mathbf{a}_p$ is the p -th row of the slope matrix $\mathbf{A}^{\text{PA}}$ in $\Omega_{\text{Approx}}^{\text{Agg}}$ (4.26). $\mathcal{X}^{\text{Agg}} \subset \mathbb{R}^T$ is a compact domain for $\mathbf{p}^{\text{Agg}}$ and can be defined by the maximum and minimum aggregated power of the network based on the installed capacity. M is a sufficiently large constant and is suggested in [108] to be

$$M = \max_{\mathbf{p}^{\text{Agg}} \in \mathcal{X}^{\text{Agg}}, p \in [P]} \{|b_{i,p}^{\text{PA}} - \mathbf{a}_p^\top \mathbf{p}^{\text{Agg}}|\}. \quad (4.34)$$

Comparing the derived $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ in Eq. (4.33) to the polytope approximation under no uncertainty in Eq. (4.26), it can be seen that the MILP constraints introduce additional ancillary terms in Eq. (4.33)(e) and (f) that tighten the original polytope constraints, which reflects the ϵ -risk-averse attitude and hedges the worst-case distribution in WDRJCC [108]. Since θ is the radius of the Wasserstein ambiguity set $\mathcal{F}_N(\theta)$, a larger θ leads to a smaller reliable AFAPR by definition.

Furthermore, it can be shown that the derived $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ in (4.33) achieves the objective in Eq. (4.25):

Proposition 3. For $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ defined in (4.33), if $\mathbb{P} \in \mathcal{F}_N(\theta)$, then $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta) \subseteq \Omega^{\text{Agg}}(\epsilon)$ in (4.24).

Proof. Note that $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ in (4.33) is equivalent to (4.32); thus for any $\mathbf{p}^{\text{Agg}} \in \Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$, it follows that

$$\mathbb{P}'(\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}) \geq 1 - \epsilon, \forall \mathbb{P}' \in \mathcal{F}_N(\theta). \quad (4.35)$$

Since $\mathbb{P} \in \mathcal{F}_N(\theta)$, it can be shown that $\mathbb{P}(\mathbf{p}^{\text{Agg}} \in \Omega_{\text{Approx}}^{\text{Agg}}) \geq 1 - \epsilon$. Combined with the definition of $\Omega_{\text{Approx}}^{\text{Agg}}(\epsilon)$ in Eq. (4.30), it follows that $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta) \subseteq \Omega_{\text{Approx}}^{\text{Agg}}(\epsilon)$.

Finally, based on Proposition 2, it can be deduced that $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta) \subseteq \Omega^{\text{Agg}}(\epsilon)$. $\square$

In other words, when the Wasserstein ambiguity set $\mathcal{F}_N(\theta)$ covers the probability function $\mathbb{P}$ for the true distribution of $\tilde{\mathbf{b}}$, Proposition 3 states that the proposed $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ (4.33) achieves the objective in (4.25). Finally, as the WDRJCC in (4.32) is exactly reformulated, the proposed inner-approximation $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ to the jointly reliable AFAPR could achieve a low conservativeness.

The low complexity of the approximation is also a part of the objective of this chapter. The reformulation in Eq. (4.33) introduces N binary variables and is composed of at most $P + P[\epsilon N] + N + 1$ constraints (exclusive of the constraints for the type of decision variables like integrity and non-negativity). The complexity of (4.33) is low due to the use of surrogate polytope JCC and the Wasserstein metric: as discussed in Section 4.2.1, for the VB polytope used in the case studies, the dimension of $\tilde{\mathbf{b}}$, i.e., P , is only $4T$, not increasing with the number of DERs nor with the size of the network. Also, the use of the Wasserstein metric leads to a good out-of-sample performance even under a small number of data samples (N). The case studies in Section 4.4 demonstrate the low complexity and the good performance for scheduling problems based on the derived $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ (4.33) with $N = 50$ and $T = 24$. Note that modelling a single generation unit needs T binaries or $3T$ binaries when the shut-down and start-up statuses are explicitly modelled [315], which is close to the $N = 50$ binaries of the proposed MILP-based AFAPR.

It is also worth noting the possible computing challenges when integrating the proposed MILP-based jointly reliable AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ into the upper-level scheduling, although the number of binaries introduced by the proposed AFAPR is comparable to a single generator. The upper-level scheduling problem may be a large-scale multi-period AC optimal power flow (ACOPF) problem to accurately model network constraints [316], which is already a computationally challenging non-linear non-convex problem. The additional binaries introduced by the proposed AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$, especially when there are several AFAPRs for several distribution networks, may amplify the issue. Furthermore, the proposed AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$

makes scheduling problem at least a MILP, which is in theory NP-hard and can be hard to solve in some bad cases. To tackle this challenge, the proposed SFLA in Chapter 3 can be applied as a further inner-approximation. As demonstrated in Chapter 3, the proposed SFLA avoids this NP issue, increases computational efficiency, and only introduces minor extra conservativeness.

Finally, the proposed MILP-based AFAPR in (4.33) can be integrated into any non-linear non-convex upper-level scheduling problems performed on a closed and bounded domain [108].

Remark 5: The proposed MILP-based AFAPR (4.33) is based on the inner polytope approximation in Section 4.2.1. One may also explore the inner ellipsoid $\{\mathbf{p}^{\text{Agg}} \mid \mathbf{p}^{\text{Agg}} = \mathbf{E}\boldsymbol{\xi} + \mathbf{e}, \|\boldsymbol{\xi}\| \leq 1\}$ to approximate the AFAPR in a deterministic setting [164]. The vector $\boldsymbol{\xi}$ collects auxiliary decision variables. Both $\mathbf{E}$ and $\mathbf{e}$ are parameters to be inferred [164] and are viewed as random variables under the proposed method to derive the reliable AFAPR as in Section 4.2.2. However, since the random $\mathbf{E}$ is multiplied with decision variables in $\boldsymbol{\xi}$, the exact analytical WDRJCC reformulation in (4.33) is not applicable. The more computationally extensive reformulation for LHS-WDRJCC [117] is needed.

Remark 6: The proposed method may tackle cases with more significant changes including network reconfiguration and the DER faults. Corresponding $\tilde{\mathbf{b}}$ samples can be generated for these cases and added to $\mathcal{D}$ in deriving the reliable AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ (4.33) to improve the robustness. These require more case study validations and are left as future work.

4.3.2 Cost and disaggregation

The derived AFAPR only characterises the feasibility region and is insufficient for economic upper-level system scheduling [317], which requires a monetisation function that maps the aggregated power to the total DER operational costs for providing flexibility. Existing work derived the cost mapping by characterising the cost zone [24, 303], in which the flexibility cost is smaller than a certain threshold, or evaluating the cost for each aggregated power sample with certain intervals [318]. However, these works did not derive the analytical cost curve, which is important to integrate efficiently into optimisation-based scheduling tools [317]. In fact, because of the complicated constraints of heterogeneous DERs and the network, the analytical cost curve is difficult to derive. Refs. [94] and [317] proposed to use a piecewise-linear approximation to derive the analytical cost curve. However, as pointed out in [319], there is still a gap in considering the intertemporality. In general, monetisation in DER aggregation remains a vital but underexplored topic. Again, the key focus in this chapter is the derivation of the jointly reliable AFPR. More accurate cost modelling is left as future work.

This chapter adopts the piecewise-linear approximation in [94] to derive the analytical approximated aggregated cost model associated with the derived AFAPR. The core idea is to fit a piece-wise function with K^{PW} pieces given $K^{\text{PW}} + 1$ pairs of the aggregated power and the minimum total DER cost for each time step, which neglects the intertemporality. The detailed

fitting procedure is provided in [94]. It should be noted that the case studies in this chapter will evaluate the proposed method using the exact low-level DER operational costs instead of the approximated aggregated cost. This chapter considers the following DER operational cost:

$$C(\mathbf{x}^{\text{DER}}) = \omega^{\text{ES}} \Delta t \sum_{t, \psi, i^{\text{ES}}} (\hat{p}_{t, i^{\text{ES}}}^{\text{ES}, \psi} - \check{p}_{t, i^{\text{ES}}}^{\text{ES}, \psi}) + \omega^{\text{CL}} \Delta t \sum_{t, \psi, i^{\text{CL}}} \max\{(\ddot{p}_{t, i^{\text{CL}}}^{\text{CL}, \psi} - p_{t, i^{\text{CL}}}^{\text{CL}, \psi}), 0\}, \quad (4.36)$$

where the first term is the charging and discharging costs of ES, and the second term models the load curtailment costs. ω^{ES} and ω^{CL} are the corresponding cost coefficients. Renewable generation is considered cost-free. The forecast mean values of uncertain DERs are used in calculating the aggregated cost model denoted as $C'(\mathbf{p}^{\text{Agg}})$, which will be submitted together with the AFAPR for upper-level system scheduling to respect the DER operation costs.

After the upper-level scheduling, the aggregator will be informed of the scheduled aggregated power $\mathbf{p}^{\text{Agg}^*}$, which is disaggregated to each DER. For each realisation of uncertain DERs, the disaggregation process can be formulated as:

$$\begin{aligned} \min_{\mathbf{x}^{\text{DER}}} \quad & C(\mathbf{x}^{\text{DER}}) + \tau^{\text{P}}(\mathbf{x}^{\text{DER}}) \\ \text{s.t.} \quad & \text{Eqs. (4.1)–(4.22)} \end{aligned} \quad (4.37)$$

where $C(\mathbf{x}^{\text{DER}})$ is the DER operation cost in (4.36), and $\tau^{\text{P}}(\mathbf{x}^{\text{DER}})$ penalises the nodal voltage estimated by the linear network model being close to the limit, which is to offset the inaccuracy of the linear network constraints. An alternative could be a re-dispatch based on the updated linear network model. This chapter does not additionally include a penalty for the ES simultaneous charging and discharging, because the modelled ES charging cost in $C(\mathbf{x}^{\text{DER}})$ has the same form of the proposed penalty term in [100, 306] explained in Section 4.1.1.

4.3.3 Overall framework of Part I

Algorithm 1 summarises the steps for using the proposed $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ with a detailed description given below. The aggregator derives the jointly reliable AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ and submits it to the upper-level system for the central scheduling. The aggregator first collects the dataset $\mathcal{D}$ that contains possible realisations of the random vector $\tilde{\mathbf{b}}$ in (4.30) in an offline process, which is then used in the scheduling stage.

Algorithm 1 Using the Proposed Jointly Reliable AFAPR**Offline Dataset Gathering Stage***Aggregator:*

- 1: **for** each historical horizon i **do**
- 2: Infer $\mathbf{b}_{\text{fore,hist},i}^{\text{PA}}$ and $\mathbf{b}_{\text{hist},i}^{\text{PA}}$ based on historical DER forecasts and true values;
- 3: Push $\mathbf{b}_{\text{error},i}^{\text{PA}} = \mathbf{b}_{\text{hist},i}^{\text{PA}} - \mathbf{b}_{\text{fore,hist},i}^{\text{PA}}$ to $\mathcal{D}_{\text{error}}$;
- 4: **end for**

Scheduling Stage*Aggregator:*

- 1: Infer $\mathbf{b}_{\text{fore}}^{\text{PA}}$ based on forecasts of uncertain DERs in the horizon to be scheduled;
- 2: Get $\mathcal{D}$ as Eq. (4.39);
- 3: Derive $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ as Eq. (4.33) and submit;

Upper-level System:

- 4: Solve scheduling problems like Eq. (4.43) and the solution will meet the desired $1 - \epsilon$ reliability level.

Offline dataset gathering

The dataset $\mathcal{D}$ should contain possible realisations of $\tilde{\mathbf{b}}^{\text{PA}}$ for the future horizon. In addition, the random vector $\tilde{\mathbf{b}}^{\text{PA}}$ is decomposed into a deterministic part $\mathbf{b}_{\text{fore}}^{\text{PA}}$ plus a random part $\tilde{\mathbf{b}}_{\text{error}}^{\text{PA}}$:

$$\tilde{\mathbf{b}}^{\text{PA}} = \mathbf{b}_{\text{fore}}^{\text{PA}} + \tilde{\mathbf{b}}_{\text{error}}^{\text{PA}}. \quad (4.38)$$

Such decomposition captures the random variable's dependency on the scheduled day by $\mathbf{b}_{\text{fore}}^{\text{PA}}$, which is commonly used in current research [38]. During the i -th historical operation horizon, the decision-maker runs a method to find $\mathbf{b}^{\text{PA}}$ (e.g., [95]) based on forecasts of uncertain DERs and the real values after the uncertainty is revealed. This thesis denotes the one based on forecasts by $\mathbf{b}_{\text{fore,hist},i}^{\text{PA}}$ and the other by $\mathbf{b}_{\text{hist},i}^{\text{PA}}$. The quantity $\mathbf{b}_{\text{error},i}^{\text{PA}} = \mathbf{b}_{\text{hist},i}^{\text{PA}} - \mathbf{b}_{\text{fore,hist},i}^{\text{PA}}$ is added to $\mathcal{D}_{\text{error}}$ as the i -th sample of the random part $\tilde{\mathbf{b}}_{\text{error}}^{\text{PA}}$.

In real-world applications, the data collection process can run parallel to the operation. A new data sample will be added to $\mathcal{D}_{\text{error}}$ every time an operation horizon ends. Additionally, the proposed method does not have stringent requirements for the size of the dataset, reducing the efforts on data collection. The case studies in Section 4.4.6 demonstrate that a $\mathcal{D}_{\text{error}}$ with data size 50 can achieve good performance for a day-ahead 24-hour scheduling problem with hourly resolution.

Scheduling stage

In the operational stage, the aggregator first performs forecasts for all the uncertain DERs for the horizon to be scheduled, which is a common step in power system scheduling problems. Then, existing methods to calculate the polytope approximation in a deterministic setting can be applied to calculate $\mathbf{b}_{\text{fore}}^{\text{PA}}$ based on the forecast values, after which each element in $\mathcal{D}_{\text{error}}$ is added to $\mathbf{b}_{\text{fore}}^{\text{PA}}$ to get the desired $\mathcal{D}$ as

$$\mathcal{D} = \{\mathbf{b}_{\text{fore}}^{\text{PA}} + \mathbf{b}_{\text{error},i}^{\text{PA}} \mid \mathbf{b}_{\text{error},i}^{\text{PA}} \in \mathcal{D}_{\text{error}}\}, \quad (4.39)$$

Then, the MILP expression (4.33) for $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ can be derived, which is submitted for upper-level system scheduling (a simple example is (4.43)). The scheduled aggregated power will be implementable without violating any DER constraints or network constraints for the entire horizon $[T]$ with joint probability $1 - \epsilon$. Note that the derived $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ is for the entire horizon $[T]$ rather than for a single time step. The aggregator can also calculate the aggregated cost associated with the AFAPR and submit them to the upper-level system; the dispatched aggregated power is finally disaggregated to individual DERs as described in Section 4.3.2.

Complexity analysis

Over the online Steps 1–3 to calculate the proposed jointly reliable AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ (4.33) in Algorithm 1, the only computationally demanding step is the derivation of $\mathbf{b}_{\text{fore}}^{\text{PA}}$ (Step 1) based on forecasts of uncertain DERs, by using existing methods that calculate the inner polytope approximation of the AFAPR in a deterministic setting (see Section 4.2.1). However, it is noteworthy that the main aim of this chapter is to improve these existing methods to have a jointly reliable probabilistic guarantee (see Remark 4). The complexity of the proposed method is on Steps 2 and 3, which involve only basic algebra operations given the collected $\tilde{\mathbf{b}}_{\text{error}}^{\text{PA}}$ samples offline. Therefore, the proposed online process to calculate the jointly reliable AFAPR leads to the same order of complexity compared to existing methods that calculate the inner polytope approximation of the AFAPR in a deterministic setting. Furthermore, as has been discussed in Section 4.2.2, as the surrogate polytope WDRJCC is defined only for the aggregated power $\mathbf{p}^{\text{Agg}}$, the computational complexity of the proposed method does not increase with the number of DERs nor with the size of the network.

The existing methods in Step 1 involve optimisation problems with complexity dependent on the number of DERs and the network sizes. However, as has been discussed in Remark 4, Algorithm 1 is capable of more efficient methods to perform Step 1. In addition, another option is to explore the data-driven probabilistic forecasting framework to generate the $\tilde{\mathbf{b}}^{\text{PA}}$ samples in $\mathcal{D}$ directly, bypassing Step 1 in calculating $\mathbf{b}_{\text{fore}}^{\text{PA}}$ and possibly leading to improved

accuracy given a sufficient amount of historical data and appropriate training. The subsequent visualisation of $\tilde{b}^{PA}$ samples in Fig. 4.7 explained in Section 4.4.3 suggests good predictability, possibly because the high stochasticity of individual DERs offset each other in reaching the aggregation parameters in b^{PA} .

4.4 Part I: case study

4.4.1 Network and DER settings

The case studies are based on the IEEE-123 distribution feeder [320] modelled in OPEN [321]. DERs include 10 ESs, 20 PVs, 20 WTs, and 30 CLs. There are also 65 NLs. The DER location and connected phases are shown in Fig. 4.4. The red colour is for phase a or ab, the green colour represents phase b or bc, and the blue colour is for the rest. The UK power demand is estimated to have 20%-40% curtailment potential by 2050 [9]. This chapter uses an intermediate and symmetric setting: it is assumed that the CLs are controllable with a $\pm 30\%$ range. The voltage limits in Eq. (4.22) are set to 0.95 to 1.05 p.u. as [94]. The DER cost coefficients ω_{iES}^{ES} and ω_{iCL}^{CL} are both set to 8p/kWh as [322]. The sum of the peak load for the whole network is about 3500 kW. To fully validate the proposed framework, this chapter considers three case studies with different renewable energy source (RES) settings: 1) LowRES with an 800 kW total PV capacity and an 800 kW WT capacity; 2) HighWT, in which the total WT capacity increases to 5600 kW; and 3) HighPV, in which the total PV capacity increases to 7000 kW. For all three case studies, the total ES size is 1800 kWh and the setting for CLs and NLs is kept unchanged. All three case studies have stressful initial network conditions. Fig. 4.5 illustrates the nodal voltages for the three case studies under an 'unregulated' profile, where RESs are fully utilised with unity power factors, CLs are maintained at the base levels, and ESs are controlled to minimise their own energy costs. Under-voltage issues occur in the 'LowRES' case due to the peak network load. Both over-voltage and under-voltage issues exist in the 'HighPV' case due to the peak solar generation at some nodes and the high demand at other nodes far away from PVs. Finally, in 'HighWT', severe over-voltage issues occur in the morning and at night due to the low demand but high wind availability.

As mentioned in Section 4.1.3, the base load for CL and NL, and the maximum PV and WT generation are considered uncertain. These uncertain parameters are modelled as deterministic base curves (forecasts) plus random forecasting errors (random variables) with hourly granularity. The base curve for RES comes from the real-world dataset in [323], and the base curve for network load is scaled from the load data for US commercial buildings [324].

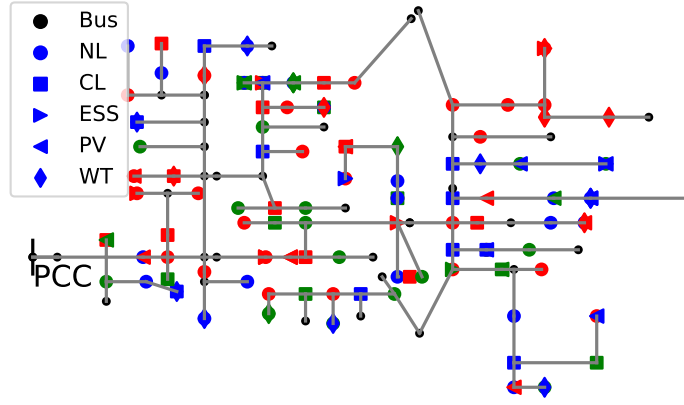


Figure 4.4: DER locations in the IEEE-123 distribution network.

The red colour highlights the connection to phase a or ab, the green colour is for phase b or bc, and the blue colour is for the remaining phases.

The RES forecasting errors are modelled based on real-world forecasting error statistics in [127, 325, 326]. For the same type of RES, the correlation of the forecasting errors in different connection sites are set to 1, considering a relatively small coverage of a distribution grid. The WT and PV error histogram is plotted in Fig. 4.6, which shows a non-Gaussian pattern (note that, this non-Gaussian error histogram represents the marginal distribution. However the one of more interest in decision making is the conditional distribution of each decision-making time step and their correlation. A non-Gaussian marginal distribution does not necessarily imply non-Gaussian conditional distributions). The forecasting error for the network load is the only one synthesised as Gaussian distribution with zero mean and a standard deviation equal to 5% of the base load curves. The magnitude of this standard deviation is close to the day-ahead commercial or industrial building load forecasting errors [327]. Ref. [328] showed that the Gaussian distribution can fit the load forecasting error well. This chapter sets the correlation of the load forecasting error among buses as 0.5. In the ‘LowRES’ case the Gaussian error of the load can be dominant, but the real-world non-Gaussian RES error data will be dominant in the ‘HighPV’ and ‘HighWT’ cases.

Based on the forecasting error modelling, the case studies generate 500 samples per time step of the scheduling horizon $[T]$ as the training set, which is for constructing the proposed jointly reliable AFAPR and other benchmarks. Note that the proposed method ($\mathcal{D}$ in the ‘Offline Dataset Gathering Stage’) is implemented based on part of the training set, e.g., only randomly picked 50 in Section 4.4.6. Furthermore, an out-of-sample test set is generated, which contains 300 samples per time step for the out-of-sample cost and reliability evaluation of different models.

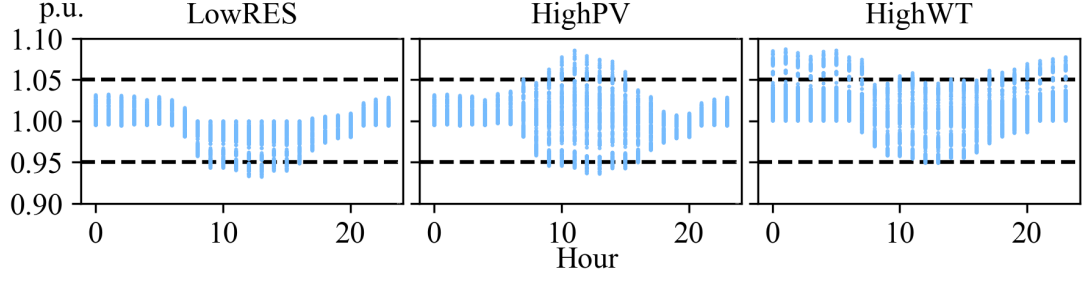


Figure 4.5: Nodal voltages for all phases at all nodes under an unregulated baseline profile. In the unregulated baseline profile, RESs are fully utilised with unity power factors, CLs are maintained at the base levels, and ESs are controlled to minimize their own energy costs. Dashed lines are the voltage limits set in the case studies.

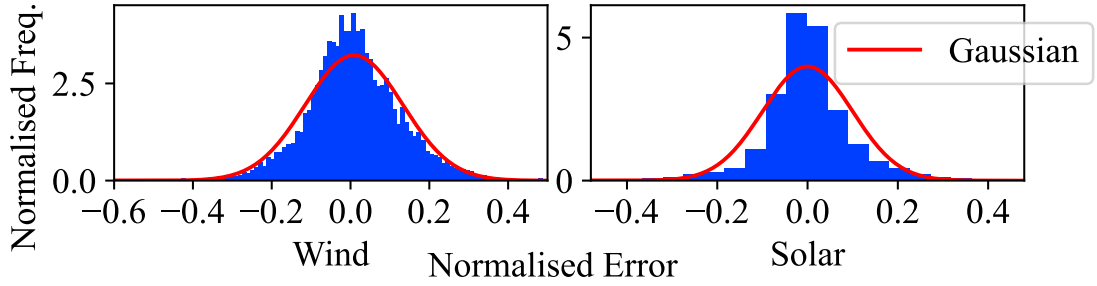


Figure 4.6: Histogram of the real-world wind and solar forecasting error.

Note that, this error histogram represents the marginal distribution. However the one of more interest in decision making is the conditional distribution of each decision-making time step and their correlation. A non-Gaussian marginal distribution does not necessarily imply non-Gaussian conditional distributions.

4.4.2 Benchmark models

The benchmark models follow existing work [94, 95, 168] that applied CC to DER aggregation. These works directly dealt with the large-scale JCC in Eq. (4.24b), unlike the proposed approach that deals with the JCC in the space of aggregated power p^{Agg} .

However, possibly due to the complicated two-stage setting and the large scale of the JCC in Eq. (4.24b), existing work did not properly tackle this JCC: they did not explicitly and formally formulate the two-stage JCC, and directly applied a single-stage CC formulation [94, 95, 168]. In this section, two categories of existing work are described and included as benchmarks.

The first set of benchmarks follows [168]. In this work, the two-stage structure is neglected, and the JCC in (4.24) is converted as a static single-stage JCC as follows:

$$\exists \mathbf{x}^{\text{DER}}, \mathbb{P}(\text{Eqs. (4.1)–(4.22)} \forall \psi, i_d, t) \geq 1 - \epsilon. \quad (4.40)$$

As can be seen, the existence for the DER power is lifted out of the JCC, which means the disaggregation $\mathbf{x}^{\text{DER}}$ is fixed (static) before the uncertainty realisation for the next T time steps, and this static $\mathbf{x}^{\text{DER}}$ can satisfy all constraints inside of the JCC (4.40) with high probability $1 - \epsilon$. This static structure can have two limitations. First, the adjustability of the DER power

x^{DER} is neglected, which can lead to over-conservativeness. Second, the equality constraint (4.20) that involves uncertain NL variables cannot be satisfied in a static JCC, which will always have a zero probability of satisfaction when the random variables follow a continuous probability distribution. To tackle this issue, Ref. [168] neglects this ill-posed equality constraint by assuming no uncertainty for NL, and then lifts all these no-uncertainty constraints out of the static JCC, which will, however, lead to reduced reliability.

Even after the processing above, the processed static JCC is still difficult to handle due to its large scale, given that the JCC includes all DER constraints and network constraints. Therefore, Ref. [168] further applied ICCs as an approximation to the static JCC (4.40). Depending on the way to set the risk level of the ICCs, this chapter further sets up two benchmarks:

1. B1: This is a safe Bonferroni approximation to the static JCC (4.40), which sets a lower risk level $\epsilon_I = \epsilon/N^J$ for each ICC, where N^J is the number of constraints in the static JCC.
2. B1unsafe: This sets all ICCs to have the same risk level ϵ as the static JCC (4.40). Although the static structure has already led to extra conservativeness, as will be demonstrated in the case studies, B1unsafe will still lead to unreliable results.

Note that, when evaluating all the benchmarks (including B2 and B2unsafe that will be introduced next), this chapter follows the same disaggregation process as the proposed method by solving Eq. (4.37). In this way, there are different disaggregation DER actions x^{DER} for each uncertainty realisation, which enables the adjustability for the static CC framework in B1 and B1unsafe. However, even under this setting, B1unsafe is still demonstrated to be unreliable.

The second set of benchmarks follows [94, 95], which processed the JCC in a way between the static and the two-stage frameworks. More specifically, compared to the processed JCC in B1 and B1unsafe, all the network inequality constraints are kept to be static, while other constraints in the JCC are replaced with the following constraints that govern the maximum/minimum range of the summed power:

$$\sum_{i_{d'}, \psi}^{N^{d'}} p_t^{d'} - \sum_{i_{\text{NL}}, \psi}^{N^{\text{NL}}} \dot{p}_t^{\text{NL,fore}} \leq \sum_{i_{d_R}, \psi}^{N^{d_R}} \bar{p}_t^{d_R} - \sum_{i_{\text{CL}}, \psi}^{N^{\text{CL}}} \sigma_{\min} \dot{p}_t^{\text{CL}} - \sum_{i_{\text{NL}}, \psi}^{N^{\text{NL}}} \dot{p}_t^{\text{NL}}, \quad (4.41)$$

$$\sum_{i_{d'}, \psi}^{N^{d'}} p_t^{d'} - \sum_{i_{\text{NL}}, \psi}^{N^{\text{NL}}} \dot{p}_t^{\text{NL,fore}} \geq \sum_{i_{\text{CL}}, \psi}^{N^{\text{CL}}} \sigma_{\max} \dot{p}_t^{\text{CL}} - \sum_{i_{\text{NL}}, \psi}^{N^{\text{NL}}} \dot{p}_t^{\text{NL}}, \quad (4.42)$$

where $d' \in [\text{PV}, \text{WT}, \text{CL}]$ and $d_R \in [\text{PV}, \text{WT}]$. The phase index ψ and bus index i are hidden for simplicity. Constraint (4.41) means the summed power scheduling of all the uncertain DERs plus the forecasted NL should be smaller than the case when the generation reaches the maximum and the load reaches the minimum (these maximum and minimum are uncertain

parameters); constraint (4.42) is for the other side. This approach is similar to the proposed surrogate polytope method that constrains p^{Agg} directly, which thus partially inherits the two-stage modelling. However, there is no guaranteed inner-approximation compared to the proposed method in Propositions 2 and 3.

The second set of benchmarks still has a complicated JCC due to the numerous number of network constraints. Therefore, Refs. [94, 95] also used the ICC framework. This chapter then constructs B2 and B2unsafe based on the two approaches to set the ICC risk levels, which is the same as the construction for B1 and B1unsafe. Finally, similar to B1 and B1unsafe, constraints that do not involve uncertain parameters are included as standard deterministic constraints in B2 and B2unsafe.

Among all benchmark models, the Gaussian distribution is used to model the uncertain parameters. This Gaussian setting is common as it enables an efficient transformation of ICCs into solvable counterparts based on the corresponding percentiles. All 500 training data are used to estimate the Gaussian distribution for different uncertain parameters in the benchmark models.

Finally, unlike the proposed method, all these benchmarks are not aggregated (projected) to the space of the aggregated power p^{Agg} . This setting gives an advantage to the benchmarks as the loss of accuracy over the aggregation process is not triggered. However, this setting cannot be applied in practice as this is equivalent to intractable direct scheduling. Note that the evaluation of the p^{Agg} -space in subsequent case studies can still be applied to these benchmarks, by solving feasibility problems to check whether there exists a feasible x^{DER} for each evaluation point in the p^{Agg} -space.

Remark 7: Note that, alternatively, one may also directly reformulate the JCC with the two-stage structure in Eq. (4.24b), and then perform aggregation algorithms to calculate the AFPR. However, due to the large scale of the JCC, this approach can be computationally intensive. Furthermore, many DER aggregation algorithms require convexity [25, 100] or even linearity [94, 95, 102], and will introduce greater computational efforts, which is incompatible with many reformulation schemes of the JCC with the two-stage structure. Even with the linear decision rule [145], the JCC will become a large-scale LHS-JCC [145], which is generally non-linear, non-convex, and computationally demanding as discussed in Section 2.2.2; thus the subsequent aggregation process is even less computationally tractable.

4.4.3 Setting of the proposed method

As the DERs in this chapter only have energy and power constraints, the VB polytope approximation is applied. The bound shrinking method in [95] is used to obtain the VB parameters as described in Section 4.2.1. Then, $\mathcal{D}$ is collected based on the process in Algorithm 1: 50 randomly selected data samples from the training set are used to generate $\mathbf{b}_{\text{hist}}^{\text{VB},i}$ with $i \in \{1, \dots, 50\}$, simulating those collected over 50 historical operation horizons. Forecasts of uncertain DERs are used to calculate $\mathbf{b}_{\text{fore}}^{\text{VB}}$ for the scheduled horizon and $\mathbf{b}_{\text{fore, hist}}^{\text{VB},i}$ for historical horizons. The subtraction between those $\mathbf{b}_{\text{hist}}^{\text{VB},i}$ and $\mathbf{b}_{\text{fore, hist}}^{\text{VB},i}$ lead to the 50 historical $\tilde{\mathbf{b}}_{\text{error}}^{\text{VB}}$ samples ($\mathcal{D}_{\text{error}}$), which, combined with $\mathbf{b}_{\text{fore}}^{\text{VB}}$, results in $\mathcal{D}$. Fig. 4.7 displays 5 of the collected $\tilde{\mathbf{b}}^{\text{VB}}$ samples in $\mathcal{D}$ for the ‘LowRES’ case. A $\tilde{\mathbf{b}}^{\text{VB}}$ sample consists of power and energy boundary parameters as (4.29). Each pair of the ‘Pmax’ and ‘Pmin’ are the power boundaries of aggregated active power that can be adjusted within for each time step, while each pair of ‘Emax’ and ‘Emin’ represents the energy bounds for the accumulated aggregated active power. Unlike a real battery, the energy boundaries of VB are derived to closely capture the accumulated energy generation and consumption of aggregated DERs; thus display an increasing pattern.

The bound shrinking method [95] requires solving a series of MILPs, which could be time-consuming when the problem scale is large. In addition, the case studies simulate the entire sample generation process (the offline stage in Algorithm 1) that is carried out offline in the real-world implementation. To reduce simulation efforts, this chapter relaxes the stopping bound and set the maximum solution time of each bound shrinking MILP to 40s in the case studies after Section 4.4.6, where there are $T = 24$ and a large number of DERs.

The radius θ in the proposed WDRJCC reformulation needs to be set prior to solving the problem. A larger θ leads to a larger ambiguity set and thus a more conservative AFAPR, so the selection criterion can be finding the smallest θ while providing a reliable solution, by 5-fold cross-validation (CV). For each fold, 20% data in $\mathcal{D}$ is held out and a solution is derived based on the remaining 80% data. The reliability score in this fold is estimated by the proportion of the held-out $\tilde{\mathbf{b}}^{\text{VB}}$ samples whose corresponding polytope $\Omega_{\text{Approx}}^{\text{Agg}}$ (4.27) contains the solution. The 20% percentile of the 5 reliability scores after the CV run for a θ candidate is its final reliability score. The smallest θ with a score no less than the desired reliability level is selected as the best radius. Finally, to reduce the effect of randomness, all the subsequent results of the proposed method are averaged over 5 random runs.

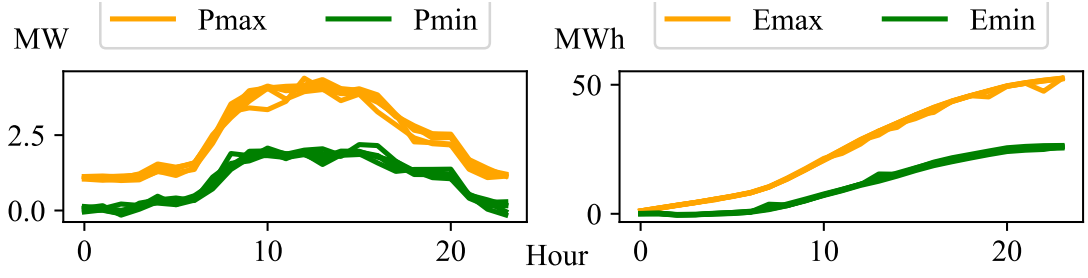


Figure 4.7: Visualisation of VB parameters of 5 of the $\tilde{b}^{VB}$ samples in $\mathcal{D}$ in the 'LowRES' case.

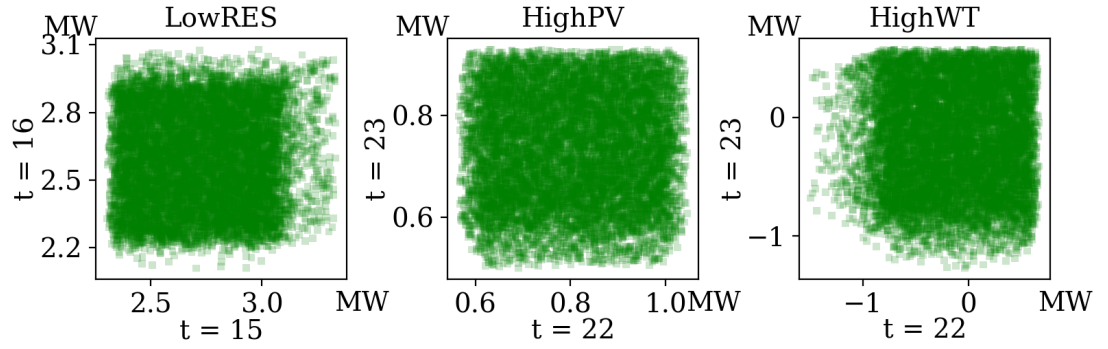


Figure 4.8: The projected 10,000 Monte-Carlo uniform samples of the proposed MILP-based AFAPR in (4.33).

4.4.4 Visualisation of the AFAPR

To facilitate understanding, 10,000 points are uniformly sampled from the proposed AFAPR in Eq. (4.33) with $T = 24$ and are visualised in Fig. 4.8. Each point corresponds to a 24-dimension vector p^{Agg} that represents the aggregated power across $T = 24$ time steps. For visualisation, these high-dimension points are projected to two adjacent time steps in Fig. 4.8. The varying density of points suggests that the proposed MILP-based AFAPR defines a region more complicated than boxes. This shape can be attributed to the VB polytope (Section 4.2.1) applied in this chapter, which involves energy constraints that cut the high-dimensional boxes defined by the power constraints. At the same time, all these projections seem to be convex, partially explaining the computational tractability of the proposed MILP-based AFAPR in Eq. (4.33).

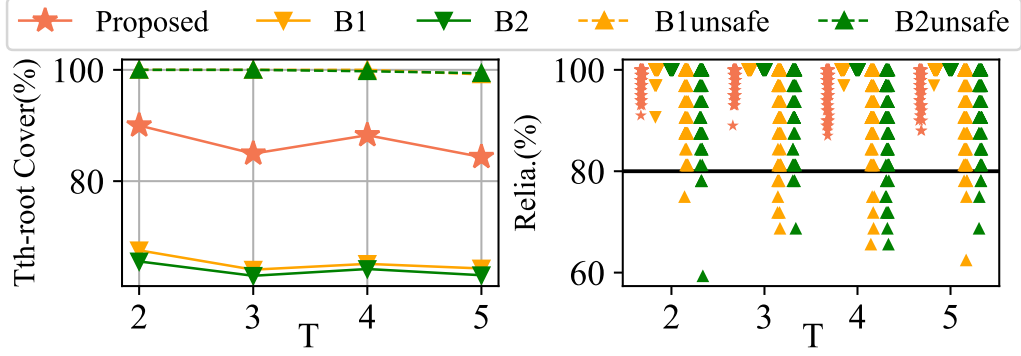


Figure 4.9: Evaluation for the whole AFAPR in conservativeness and inner-approximation. The black horizontal line is the desired reliability level $1 - \epsilon$ (80% here). An inner-approximation should have all points with reliability scores above the desired $1 - \epsilon$ level, namely no “unsafe points” with reliability below $1 - \epsilon$.

4.4.5 Evaluation of the whole AFAPR

The objective (4.25) of this chapter includes finding an inner-approximation to $\Omega^{\text{Agg}}(\epsilon)$ with low conservativeness. Therefore, this section evaluates the conservativeness and whether the inner-approximation is achieved from the perspective of the whole AFAPR. To do so, this section first uses the Monte Carlo method to uniformly sample 100 points from $\Omega^{\text{Agg}}(\epsilon)$ (constructed by out-of-sample DER scenarios) and 100 points from the proposed AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta)$ in (4.33) and the other four benchmarks. Each point corresponds to a possible $\mathbf{p}^{\text{Agg}}$, i.e., the aggregated power across the scheduling horizon $[T]$. This chapter then defines two metrics: 1) “Coverage”: the proportion of samples from the true $\Omega^{\text{Agg}}(\epsilon)$ that are covered by the proposed method and benchmarks. A higher coverage indicates a lower conservativeness; 2) “Reliability”: for each sampled point $\mathbf{p}^{\text{Agg}}$ from the proposed AFAPR and the four benchmarks, it will have a reliability score, which is the proportion of 300 out-of-sample DER scenarios that can have the sampled $\mathbf{p}^{\text{Agg}}$ disaggregated to satisfy all DER and network constraints. If the reliability score of a sampled point is above the desired level $1 - \epsilon$, it is considered a “safe point” that is inside the inner-approximation $\Omega^{\text{Agg}}(\epsilon)$; otherwise, it is an “unsafe point” outside $\Omega^{\text{Agg}}(\epsilon)$. An inner approximation should have all sampled points as “safe points”.

The applied linear network model may have limited accuracy, especially because the AFAPR represents the power range under various operating conditions. Therefore, the Z-Bus nonlinear power flow simulation [329] is used to check the violation of network constraints over the sampling from $\Omega^{\text{Agg}}(\epsilon)$ and the evaluation of “Reliability”. Results under ‘HighWT’ and $\epsilon=0.2$ for various T are given in Fig. 4.9. Note that the coverage (corresponding to volume) changes exponentially with the increase of T , so a decreasing trend can still be observed even when there is a high approximation accuracy for each dimension. Therefore, this section plots the T -th root of the coverage. It can be seen that the proposed method has a T -th root of the coverage rate around 90% for all T , significantly higher than B1 and B2. In addition, from the

right subplot of the figure, it can be seen that all samples from the proposed method have reliability scores above the desired 80% (all are “safe points”), meaning that the proposed method achieves an inner-approximation to $\Omega^{\text{Agg}}(\epsilon)$ with low conservativeness. Although B1unsafe and B2unsafe lead to around 100% coverage, they cover less reliable $\mathbf{p}^{\text{Agg}}$ samples, i.e., “unsafe points” outside $\Omega^{\text{Agg}}(\epsilon)$.

4.4.6 Scheduling evaluation

An alternative metric

In many power system scheduling problems, the scheduling horizon can be long, e.g., 24 time steps with an hourly resolution for day-ahead (DA) scheduling. The evaluation method in the previous section cannot be used for long time horizons due to the excessive complexity of Monte Carlo sampling. Note that the reliable AFAPR is used for scheduling problems, and a better AFAPR approximation tends to yield scheduling solutions with better optimality (less conservative) while still ensuring the desired solution reliability. Therefore, to assess the approximation quality, it suffices to evaluate the performance of the optimal solution $\mathbf{p}^{\text{Agg}^*}$ to the following DA cost minimisation problem with hourly resolution and $T = 24$ under the UK DA market price π^{DA} :

$$\begin{aligned} \mathbf{p}^{\text{Agg}^*} = \arg \min \quad & \pi^{\text{DA}\top} \mathbf{p}^{\text{Agg}} + C'(\mathbf{p}^{\text{Agg}}) \\ \text{s.t.} \quad & \mathbf{p}^{\text{Agg}} \in \Omega_{\text{WDR}}^{\text{Agg}}(\epsilon, \theta) \text{ in (4.33)} \end{aligned} \quad (4.43)$$

where $C'(\mathbf{p}^{\text{Agg}})$ is the aggregated DER cost derived by a piecewise-linear method in Section 4.3.2. Based on [25], one can also co-optimize the profit on reserve provision by adding $\bar{\mathbf{r}}^\top (\bar{\mathbf{p}}^{\text{Agg}} - \mathbf{p}^{\text{Agg}}) + \underline{\mathbf{r}}^\top (\mathbf{p}^{\text{Agg}} - \underline{\mathbf{p}}^{\text{Agg}})$ to the objective and $\underline{\mathbf{p}}^{\text{Agg}}, \bar{\mathbf{p}}^{\text{Agg}} \in \Omega_{\text{WDR}}^{\text{Agg}}$ to the constraints. $\bar{\mathbf{r}}$ and $\underline{\mathbf{r}}$ are the reward coefficients for the upward and downward flexibility reserve for $[T]$. $\bar{\mathbf{p}}^{\text{Agg}}$ and $\underline{\mathbf{p}}^{\text{Agg}}$ are the maximum and minimum available power defining the size of the reserve.

This section evaluates the performance of the optimal DA scheduling $\mathbf{p}^{\text{Agg}^*}$ from two perspectives: 1) Out-of-sample reliability: The disaggregation problem (4.37) is performed under each of the 300 out-of-sample DER scenarios. The nonlinear Z-Bus power flow simulation is implemented to check the network constraints of the disaggregated results. The out-of-sample reliability is defined as the proportion of the 300 out-of-sample DER scenarios where there exist successful disaggregation such that no DER constraints nor network constraints are violated. As the scheduling horizon is set to the next whole day, 300 samples correspond to roughly one-year running. The DA scheduling $\mathbf{p}^{\text{Agg}^*}$ is considered reliable as long as the out-of-sample reliability is greater than or equal to the desired level $1 - \epsilon$. 2) Out-of-sample cost: After solving the disaggregation problem (4.37) under each of the 300 out-of-sample DER scenarios, the actual cost can be obtained as $\pi^{\text{DA}\top} \mathbf{p}^{\text{Agg}^*} + C(\mathbf{x}^{\text{DER}})$. The out-of-sample cost is the average of these actual costs over the 300 out-of-sample scenarios.

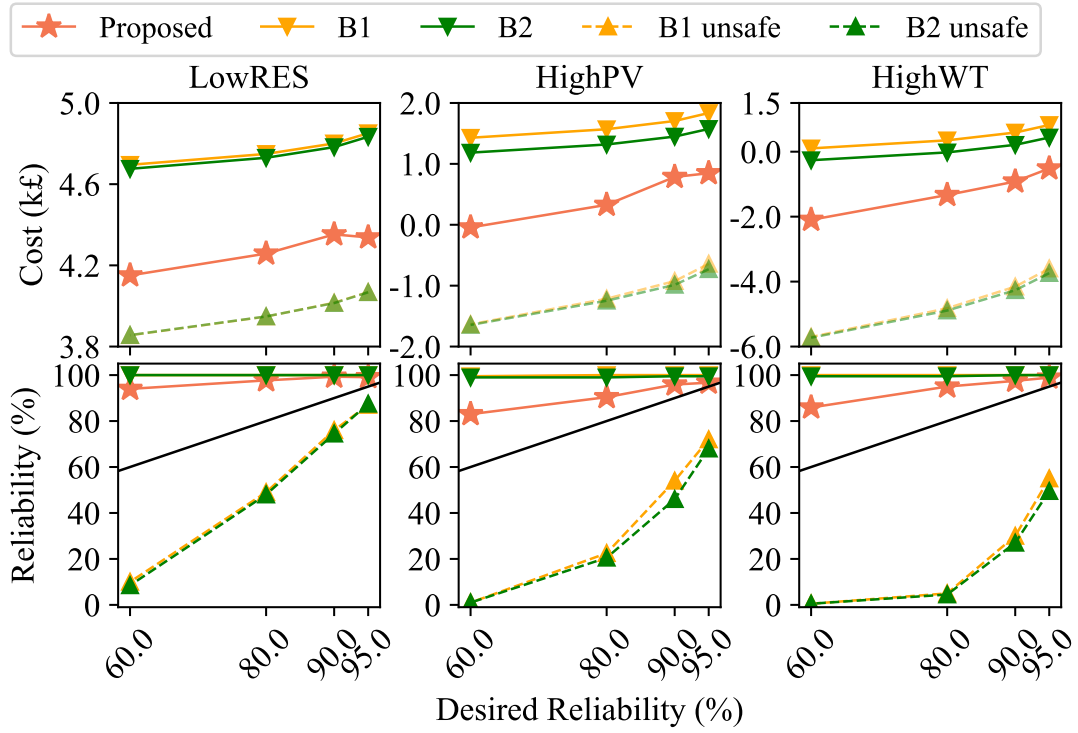


Figure 4.10: Evaluation for minimised cost (optimality) and the out-of-sample reliability.

First row: minimised day-ahead costs; second row: test reliability, where the black diagonal line represents the desired reliability levels $1 - \epsilon$.

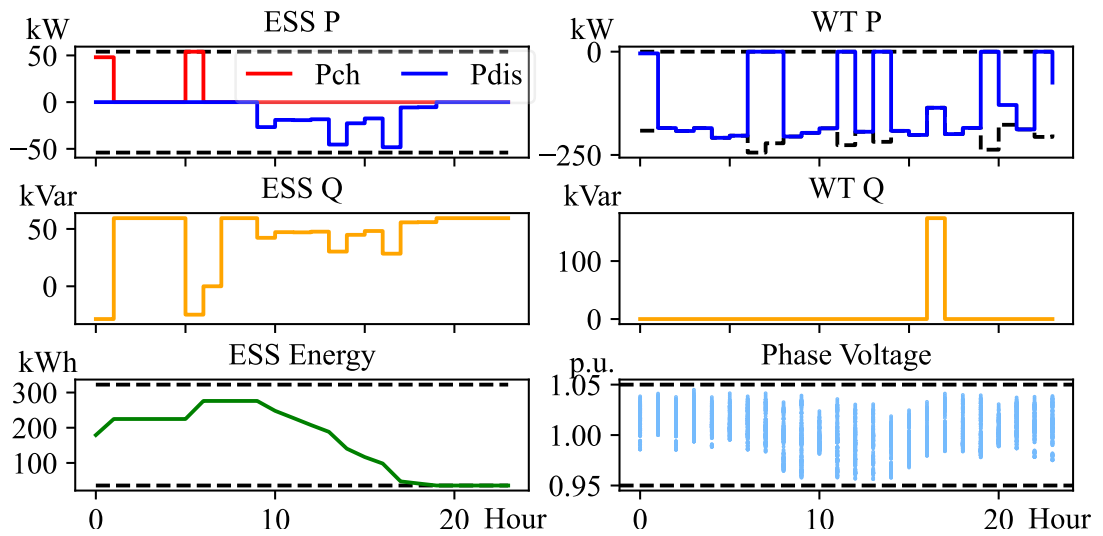


Figure 4.11: Successfully disaggregated DER power and nodal voltages (by Z-Bus nonlinear power flow) for the next $T=24$ scheduling horizon.

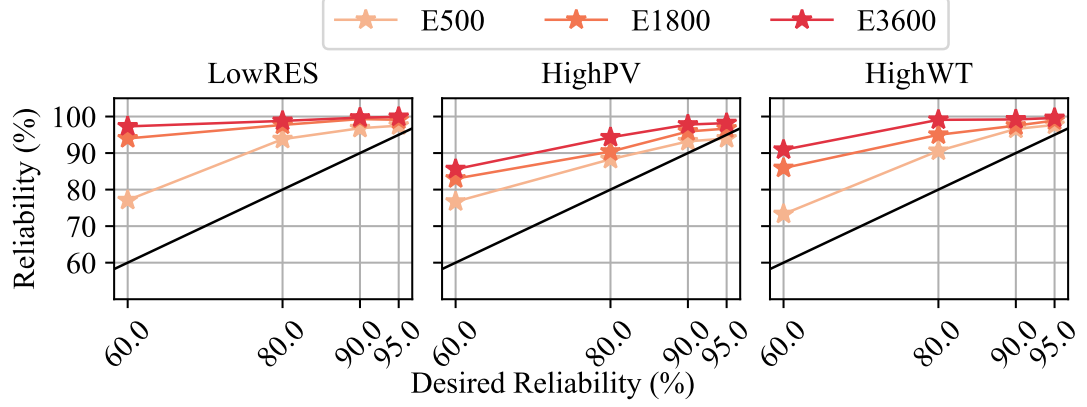


Figure 4.12: Sensitivity analysis for the out-of-sample reliability of the proposed method for different ES sizes.

The total ES sizes being evaluated include 500kWh (E500), 1800kWh (E1800), and 3600kWh (E3600). The black diagonal lines represent the desired reliability levels $1 - \epsilon$.

Results

This section plots the out-of-sample cost and out-of-sample reliability results under four desired reliability levels ($1 - \epsilon$) in Fig. 4.10. For all three RES settings and all the desired reliability levels, the proposed method achieves lower costs than B1 and B2, and the proposed method meets the reliability requirement. On the other hand, B1unsafe and B2unsafe have lower costs than the proposed method but fail to meet reliability requirements for all cases. Overall, it can be seen that the proposed method leads to less conservativeness while meeting the desired reliability.

This section also plots a successful disaggregation result, i.e., the DER trajectories across $[T]$ reaching the scheduled p^{Agg^*} , in Fig. 4.11: there is no simultaneous charging and discharging power for ES. Also, all active power, reactive power, and network voltages (by the Z-Bus power flow) are within the limits.

4.4.7 Sensitivity analysis

Size of storage

The time-coupling DERs increase the approximation difficulties of $\Omega_{\text{Approx}}^{\text{Agg}}$ [95], which may affect the performance of the proposed method. Therefore, this section evaluates the proposed method under three ES sizes: total size being 500 kWh (E500), 1800 kWh (E1800), and 3600 kWh (E3600). The maximum stored energy in the last setting can roughly support the evaluated IEEE-123 grid for one hour under the peak load time, so it is considered a large ES setting.

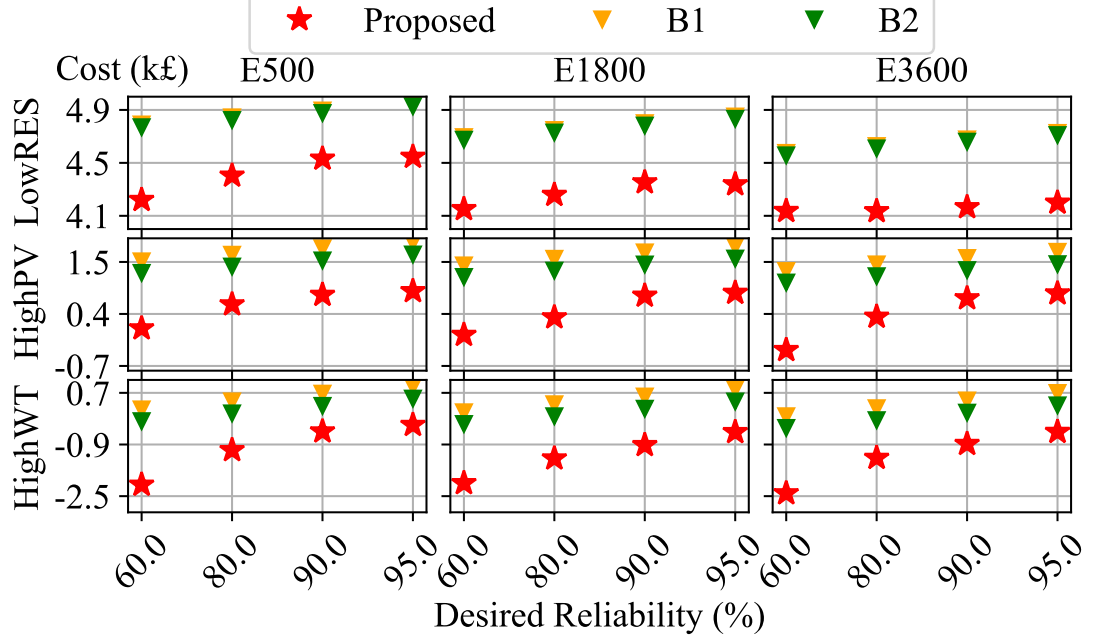


Figure 4.13: Sensitivity analysis for the out-of-sample costs of the proposed method for different ES sizes.

The total ES sizes being evaluated include 500kWh (E500), 1800kWh (E1800), and 3600kWh (E3600).

The out-of-sample reliability results are given in Fig. 4.12, which indicates that the proposed method is reliable under all ES sizes. Note that this chapter uses the inner polytope approximation, so a larger ES size leads to a more conservative approximation, thus increasing the test reliability. However, this increased conservativeness is minor compared to the reduced conservativeness by exactly reformulating the JCC in (4.33): Fig. 4.13 displays the out-of-sample costs of the proposed method, B1, and B2 when increasing the ES size. Results for B1unsafe and B2unsafe are omitted as they are unreliable. It can be seen that the proposed method is still superior for all ES sizes.

Size of $\mathcal{D}$

This section evaluates the impact of the number of $\tilde{\mathbf{b}}^{\text{VB}}$ samples in $\mathcal{D}$, i.e., the size of $\mathcal{D}$ or equivalently the number of binaries N of the proposed MILP-based AFAPR (4.33). Fig. 4.14 shows the out-of-sample cost, reliability, and the computing time of the scheduling problem (4.43) for the 'HighWT' case when the desired reliability is 80%. The costs of the reliable B1 and B2 are plotted for comparison. It can be seen that an overly small number of $\tilde{\mathbf{b}}^{\text{VB}}$ samples, i.e., 5, can lead to low cost but can be less reliable as shown in the reliability plot. Increasing

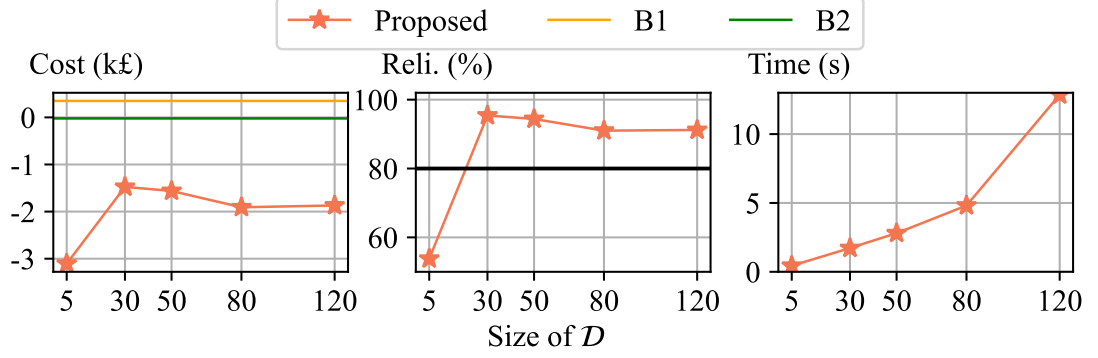


Figure 4.14: Sensitivity analysis for the out-of-sample costs and reliability of the proposed method for different number of samples.

This shows results for an increasing number of $\tilde{\mathbf{b}}^{\text{VB}}$ samples in $\mathcal{D}$, i.e., the size of $\mathcal{D}$ or equivalently the number of binaries N of the proposed MILP-based AFAPR. The dark line represents the desired reliability.

the number of $\tilde{\mathbf{b}}^{\text{VB}}$ samples from 80 to 120 results in no cost reduction (even a 2% increase) but a $1.7\times$ increase in computing time. Overall, based on Fig. 4.14, setting the number of $\tilde{\mathbf{b}}^{\text{VB}}$ samples (the number of binaries N) to within 30-80 may be more practical in use. The same conclusion can be observed in both ‘HighPV’ and ‘LowRES’ cases as well.

Note that, Ref. [111] provided a priori lower-bound estimate for the probability of the ambiguity set $\mathcal{F}_N(\theta)$ covering the true distribution. This lower-bound estimate can be calculated as a function of N (the size of $\mathcal{D}$) and the Wasserstein radius θ , which thus serves a priori guidance for choosing N and θ . However, in practice, setting N and θ based on the theoretical result can lead to over-conservativeness. It is still suggested to use the held-out method or the cross-validation method (see Section 4.4.3) to make a prudent setting [111].

There is also a trade-off in choosing the polytope approximation in (4.26). A polytope with more edges, i.e., a larger $\text{Dim}(\mathbf{b}) = P$, can have a higher approximation accuracy to the true AFAPR but also lead to more constraints in the proposed MILP-based AFAPR 4.33. A sensitivity analysis for the performance versus the computing time can be performed to select the best polytope for specific applications. In general, the selected VB approximation has a good performance and low complexity for DERs with only power and energy constraints, which has been demonstrated in the case studies of this chapter and other related literature [95].

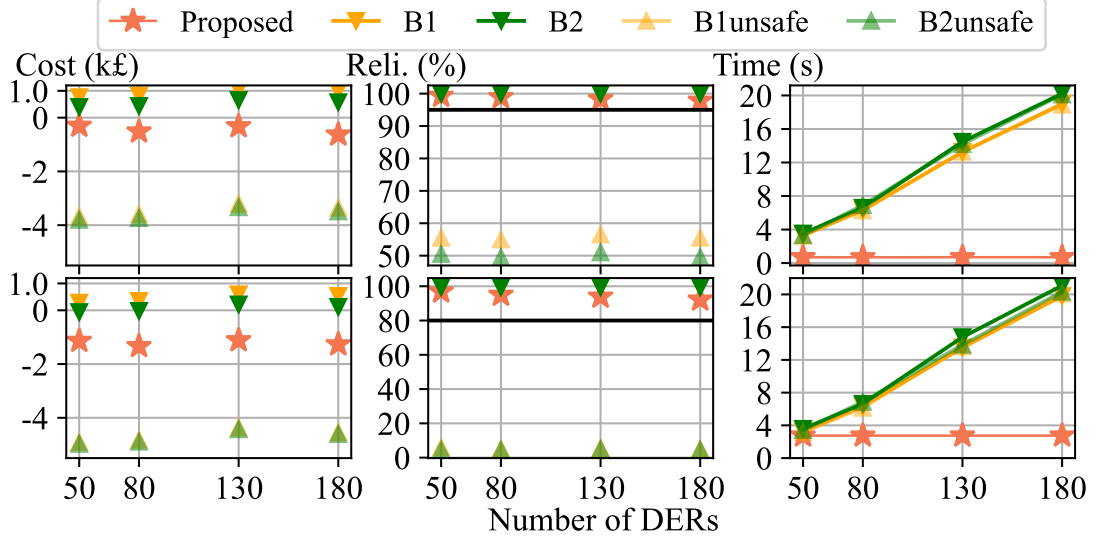


Figure 4.15: Out-of-sample costs, reliability, and scheduling time under various DER numbers.

The black horizontal line is the desired reliability level $1 - \epsilon$.

4.4.8 Scalability test

This section tests the model performance under varying numbers of DERs from 50 to 180. This section carries out the same cost-reliability analysis for ‘highWT’ under $\epsilon = 0.05$ (first row) and $\epsilon = 0.2$ (second row) in Fig. 4.15. The total capacity for each type of DERs is kept consistent for a better comparison. For the out-of-sample cost in the first column, the proposed method consistently outperforms B1 and B2. Although B1unsafe and B2unsafe display lower costs, they fail to meet the desired reliability lines (dark horizontal lines) in the middle-column plot. Conversely, the proposed method consistently yields reliable solutions, suggesting that the proposed method retains its superiority across different problem scales. This section also plots the scheduling time, namely the time of solving the scheduling problem (4.43), using the proposed method and the benchmarks in the right-most column of Fig. 4.15. It can be seen that the complexity of the proposed method is not affected by the number of DERs and the scheduling time is within seconds. The results suggest that the proposed jointly reliable AFAPR $\Omega_{\text{WDR}}^{\text{Agg}}(\epsilon)$ (4.33) can lead to computationally efficient upper-level scheduling problems.

4.4.9 Computing time of the AFAPR calculation

This section analyses the computing time of the whole online process used to calculate the reliable AFAPR (4.33) as shown in Algorithm 1. As discussed in Section 4.3.3, the only computationally demanding step is the derivation of $\mathbf{b}_{\text{fore}}$ (Step 1) based on forecasts of uncertain DERs, by using existing methods that calculate the inner polytope approximation of the AFAPR in a deterministic setting; other steps involve only simple algebra operations given the collected $\tilde{\mathbf{b}}_{\text{error}}$ samples offline and are computationally trivial. The bound shrinking method in [95] is applied to obtain $\mathbf{b}_{\text{fore}}$, which involves solving a series of MILPs. Based on the relaxed MILP criterion described in Section 4.4.3, the total computing time in Step 1 is 150 to 300 seconds for DER numbers of 50 to 180 with an AMD Ryzen 9 5900X CPU. Note that Algorithm 1 is capable of more efficient methods to perform Step 1; a possible choice is Ref. [330] that proposed a decomposition-based method that significantly accelerated the process in obtaining the inner polytope approximation with better performance than [95]. More importantly, as analysed in Section 4.3.3, the proposed online process to calculate the jointly reliable AFAPR does not change the order of complexity compared with existing methods used to calculate the inner polytope approximation of the AFAPR in a deterministic setting.

As discussed in Section 4.3.2, it is important to derive the aggregated cost function. The piecewise-linear method [94] applied to derive the aggregate cost mapping involves solving $T \times N^P$ independent (parallelisable) linear programmes, where $N^P = 5$ is the number of segments in the piecewise-linear function. Each linear programme only takes 1.5-4 seconds for DER numbers of 50 to 180, which is computationally tractable.

4.5 Part II: integrating classical JCC

The second part of the chapter shows that the proposed method can lead to tractable integration of classical JCC.

4.5.1 Proposed solution to joint chance constraint

The “surrogate polytope JCC” $\mathbb{P}(\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}} \leq \tilde{\mathbf{b}}^{\text{PA}}) \geq 1 - \epsilon$ in Eq. (4.30) can be directly tackled in the classical JCC framework. To show this, first off, the classical JCC can be converted into the following form:

$$\mathbb{P}(\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}} \leq \tilde{\mathbf{b}}^{\text{PA}}) \geq 1 - \epsilon \Leftrightarrow F_{-\tilde{\mathbf{b}}^{\text{PA}}}(-\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}}) \geq 1 - \epsilon, \quad (4.44)$$

where $F(\cdot)$ is the cumulative distribution function (CDF) for the random variable $-\mathbf{b}$.

The random vector $\tilde{\mathbf{b}}^{\text{PA}}$ can be fit by a multivariate Gaussian distribution; then based on Theorem 10.2.1 in [105], the CDF $F_{-\tilde{\mathbf{b}}^{\text{PA}}}$ is a log-concave function (and also quasiconcave) and subsequently the surrogate JCC (4.44) is convex based on the definition of quasiconcave function in Chapter 3.4.1 of [26].

Unfortunately, unlike the WDRJCC case in Section 4.3 where an analytical MILP reformulation can be derived, the classical JCC case in this chapter does not admit analytical reformulation, even though it is convex.

However, from a system operator's perspective, the objective is to calculate the optimal scheduling of the aggregated power $\mathbf{p}^{\text{Agg}}$ that is jointly reliable with the greatest economic benefit, rather than the whole feasible set of the jointly reliable $\mathbf{p}^{\text{Agg}}$. Fortunately, it is possible to calculate the desired optimal dispatch directly, through an iterative supporting hyperplane algorithm [105].

The upper-level system operator's scheduling problem can be defined as below:

$$\min_{\mathbf{p}^{\text{SO}} \in \mathcal{X}^{\text{SO}}, \mathbf{p}^{\text{Agg}} \in \mathcal{X}} \mathcal{J}(\mathbf{p}^{\text{SO}}) + C'(\mathbf{p}^{\text{Agg}}) \quad (4.45a)$$

$$\text{s.t.} \quad \mathbb{P}(\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}} \leq \tilde{\mathbf{b}}^{\text{PA}}) \geq 1 - \epsilon, \quad (4.45b)$$

where $\mathbf{p}^{\text{SO}}$ represents the upper-level system operator's other decision variables with a cost function $\mathcal{J}(\cdot)$, and $\mathcal{X}^{\text{SO}}$ is the corresponding feasible set. $C'(\cdot)$ is the cost function of the aggregated power, which is the same as the proposed Part I method in Section 4.3.2. The proposed method in Part II restricts $\mathcal{X}^{\text{SO}}$ to be convex for the validity of the supporting hyperplane algorithm [105] (note that the proposed WDRJCC method in Section 4.3 does not require this restriction). This is a reasonable assumption under the intensively used DCOPF scheme in power system ED problems. It is also possible to apply the extended supporting hyperplane algorithm proposed by [132], such that $\mathcal{X}^{\text{SO}}$ can be relaxed to MIP as the UC case. The flowchart in Fig. 4.1 needs to be updated slightly: instead of submitting the AFAPR, the aggregator now submits information sufficient for the upper-level system operator to solve Eq. (4.45) using the supporting hyperplane algorithm to get the optimal jointly reliable dispatch of the aggregated power $\mathbf{p}^{\text{Agg}}$. The supporting hyperplane algorithm is described in detail as below:

Supporting hyperplane algorithm

The core idea of the supporting hyperplane algorithm is to replace the surrogate polytope JCC in an abstract form (4.44) with a set of iteratively tightened linear constraints denoted by $\mathcal{K}^j$, with j denoting the index of the iteration. The iteration process ends when the scheduling problem based on $\mathcal{K}^j$ leads to a solution $\mathbf{p}^{\text{Agg}^*}$ satisfying $F_{-\tilde{\mathbf{b}}^{\text{PA}}}(-\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}^*}) \geq 1 - \epsilon$.

At the start of the algorithm, $\mathcal{K}^0$ needs to be initialised to be large enough to cover the convex region defined by the JCC (4.44). One option is defining a set of ICCs of the polytope constraints $\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}} \leq \mathbf{b}^{\text{PA}}$ with the same probability level as the JCC:

$$\mathcal{K}^0 = \{\mathbb{P}_{\tilde{b}_i}(\mathbf{a}_i^{\text{PA}\top} \mathbf{p}^{\text{Agg}} \leq \tilde{b}_i^{\text{PA}}) \geq 1 - \epsilon, \forall p \in [P]\} \quad (4.46)$$

$$= \{\mathbf{a}_p^{\text{PA}\top} \mathbf{p}^{\text{Agg}} \leq F_{\tilde{b}_i^{\text{PA}}}^{-1}(\epsilon), \forall p \in [P]\}, \quad (4.47)$$

where $\mathbf{a}_p^{\text{PA}}$ is the p -th row of $\mathbf{A}^{\text{PA}}$, and $\tilde{b}_p^{\text{PA}}$ is the p -th element of the random vector $\tilde{\mathbf{b}}^{\text{PA}}$. $F_{\tilde{b}_i^{\text{PA}}}^{-1}(\epsilon)$ is the inverse CDF (percentile function) of a one-dimension random variable $\tilde{b}_p^{\text{PA}}$ that can be trivially evaluated. In addition, the initialisation involves finding a feasible solution $\mathbf{p}^{\text{Agg},0}$ to the JCC (4.44), which can be obtained by solving the following feasibility problem constrained by the conservative Bonferroni approximation to the surrogate JCC (4.44):

$$\min_{\mathbf{p}^{\text{SO}} \in \mathcal{X}^{\text{SO}}, \mathbf{p}^{\text{Agg}} \in \mathcal{X}} 0 \quad (4.48a)$$

$$\text{s.t.} \quad \mathbf{a}_p^{\text{PA}\top} \mathbf{p}^{\text{Agg}} \leq F_{\tilde{b}_p^{\text{PA}}}^{-1}(\epsilon/P), \forall p \in [P]. \quad (4.48b)$$

An iterative process starts after initialisation. Over the j -th iteration with $j \geq 0$, the following optimization problem is solved:

$$\min_{\mathbf{p}^{\text{SO}} \in \mathcal{X}^{\text{SO}}, \mathbf{p}^{\text{Agg}} \in \mathcal{X}} \mathcal{J}(\mathbf{p}^{\text{SO}}) + C'(\mathbf{p}^{\text{Agg}}) \quad (4.49a)$$

$$\text{s.t.} \quad \mathcal{K}^j, \quad (4.49b)$$

where C' is the aggregated DER cost function as has been introduced in Section 4.3.2. Denote the optimal solution to (4.49) by $\mathbf{p}^{\text{Agg}^*,j}$; then, constraint (4.44) can be evaluated by calculating the CDF $F_{-\mathbf{b}}(-\mathbf{A} \mathbf{p}^{\text{Agg}^*,j})$. If constraint (4.44) is satisfied, then the problem is solved; otherwise, the algorithm finds a new boundary solution $\mathbf{p}^{\text{Agg},\text{B},j}$ between the optimal solution $\mathbf{p}^{\text{Agg}^*,j}$ over the j -th iteration and the initialised feasible solution $\mathbf{p}^{\text{Agg},0}$:

$$\mathbf{p}^{\text{Agg},\text{B},j} = \lambda \mathbf{p}^{\text{Agg}^*,j} + (1 - \lambda) \mathbf{p}^{\text{Agg},0}, \quad (4.50)$$

such that $F_{-\tilde{\mathbf{b}}}(-\mathbf{A}\mathbf{p}^{\text{Agg,B},j}) = 1 - \epsilon$. The value of $\lambda \in [0, 1]$ can be obtained by bisection [105]. Note that the boundary point $\mathbf{p}^{\text{Agg,B},j}$ is a feasible solution to the JCC (4.44) but not necessarily an optimal solution. Based on this boundary point, a new linear constraint representing the hyperplane tangent to the JCC boundary at $\mathbf{p}^{\text{Agg,B},j}$ is constructed and added to $\mathcal{K}^j$. The new constraint is:

$$\nabla_{\mathbf{p}^{\text{Agg}}} F_{-\tilde{\mathbf{b}}}^{\text{PA}}(-\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg,B},j})^\top (\mathbf{p}^{\text{Agg}} - \mathbf{p}^{\text{Agg,B},j}) \geq 0, \quad (4.51)$$

where the gradient of the multivariate Gaussian CDF can be calculated based on the Lemma 2.7.5 in [331].

The above procedure has a convergence guarantee [105], which means there will finally be one iteration such that the solution $\mathbf{p}^{\text{Agg}*,j}$ satisfies the surrogate JCC constraint (4.44). Suppose this is the W -th iteration, and the final scheduling problem with the proposed surrogate polytope JCC (4.44) can be expressed as the following problem with a set of linear constraints that substitutes the surrogate-polytope JCC:

$$\min_{\mathbf{p}^{\text{SO}} \in \mathcal{X}^{\text{SO}}, \mathbf{p}^{\text{Agg}} \in \mathcal{X}} \mathcal{J}(\mathbf{p}^{\text{SO}}) + C'(\mathbf{p}^{\text{Agg}}) \quad (4.52a)$$

$$\text{s.t.} \quad \mathcal{K}^W \quad (4.52b)$$

with

$$\mathcal{K}^W = \left\{ \begin{array}{l} \mathbf{a}_i^{\text{PA}\top} \mathbf{p}^{\text{Agg}} \leq F_{\tilde{\mathbf{b}}_i^{\text{PA}}}^{-1}(\epsilon), \forall i \in [P], \\ \nabla_{\mathbf{p}^{\text{Agg}}} F_{-\tilde{\mathbf{b}}}^{\text{PA}}(-\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg,B},j})^\top (\mathbf{p}^{\text{Agg}} - \mathbf{p}^{\text{Agg,B},j}) \geq 0, \\ \forall j \in [W] \end{array} \right\}. \quad (4.53)$$

The optimal solution $\mathbf{p}^{\text{Agg}*,j}$ with $j = W$ to (4.52) is the DA scheduled power with a $1 - \epsilon$ joint probabilistic guarantee. The whole process of the supporting hyperplane algorithm is summarised in Algorithm 2.

Remark 8: Based on Theorem 10.2.1 in [105], the convexity of the proposed surrogate JCC (4.44) holds for all probability distributions with log-concave density, including skewed distributions like Weibull, multivariate Gamma, and Dirichlet (the multivariate generalisation of the Beta) distributions with shape parameters greater than or equal to 1, and the heavily tailed hyperbolic secant distribution. These distributions are applicable for modelling a wide range of uncertain variables. However, only a subset of log-concave functions, like multivariate Gaussian, Gamma, and Dirichlet, have efficient gradient calculation when applying the supporting hyperplane algorithm [331]. Other solution algorithms, such as [332] that avoids the gradient evaluation of the CDF, may be investigated. Gaussian mixtures under certain conditions may also be applied, as they can preserve log-concavity and their CDF is the weighted summation

Algorithm 2 Supporting hyperplane algorithm

```

1: Initialise  $\mathcal{K}^j \leftarrow \mathcal{K}^0$  as (4.47);
2: Initialise the feasible solution  $\mathbf{p}^{\text{Agg},0}$  by solving (4.48);
3: Initialise  $j \leftarrow 0$ ;
4: while True do
5:   Get the  $j$ -th optimal solution  $\mathbf{p}^{\text{Agg}*,j}$  by solving (4.49);
6:   if  $F_{-\tilde{\mathbf{b}}^{\text{PA}}}(-\mathbf{A}^{\text{PA}} \mathbf{p}^{\text{Agg}*,j}) \geq 1 - \epsilon$  then
7:     Stop; output  $\mathbf{p}^{\text{Agg}*} \leftarrow \mathbf{p}^{\text{Agg}*,j}$ ;
8:   end if
9:   Find the boundary solution  $\mathbf{p}^{\text{Agg},B,j}$  (4.50) by bisection;
10:  Append the new linear constraint (supporting hyperplane) (4.51) to  $\mathcal{K}^j$ ;
11:   $j \leftarrow j + 1$ ;
12: end while

```

of Gaussian CDFs; thus the gradient can be obtained. Note that although $\tilde{\mathbf{b}}^{\text{PA}}$ is fit by a multivariate Gaussian distribution, the low-level random variables, i.e., DER power capability, do not necessarily need to follow a Gaussian distribution. The case studies of this chapter will verify this by using non-Gaussian real-world wind and solar scenarios.

Remark 9: The convexity of (4.44) and the gradient calculation in the supporting hyperplane algorithm require the distribution of $\tilde{\mathbf{b}}^{\text{PA}}$ to be non-degenerate. Non-degeneracy can be guaranteed by adding $\sigma \mathbf{I}$ to the covariance matrix, where σ is a small positive number and $\mathbf{I}$ is the identity matrix. Since all covariance matrices are positive semi-definite, σ can be set arbitrarily small to minimize the disturbance on the distribution while maintaining the positive definite property of the covariance, so as the non-degeneracy of the distribution.

4.5.2 Overall framework of Part II

The overall framework to integrate classical JCC into DER aggregation is summarised in Algorithm 3, which follows the same structure as Algorithm 1. The same offline stage is performed to collect the samples of $\tilde{\mathbf{b}}_{\text{error}}^{\text{PA}}$. In the online scheduling stage, the full samples for $\tilde{\mathbf{b}}^{\text{PA}}$ can be created and are submitted to the upper-level system operator as discussed in Section 4.5.1, who will perform the supporting hyperplane algorithm in Algorithm 2 to solve the upper-level system scheduling problem in Eq. (4.45). The solution of the aggregated power will be jointly reliable at the desired level $1 - \epsilon$. Note that, in contrast to Part I: integrating WDRJCC, the integration of classical JCC in this part can be less desirable because there is no analytical expression that can be submitted to the upper-level system operator, which may compromise computational tractability.

Algorithm 3 Using the Proposed Jointly Reliable AFAPR**Offline Dataset Gathering Stage***Aggregator:*

- 1: **for** each historical horizon i **do**
- 2: Infer $\mathbf{b}_{\text{fore,hist},i}^{\text{PA}}$ and $\mathbf{b}_{\text{hist},i}^{\text{PA}}$ based on historical DER forecasts and true values;
- 3: Push $\mathbf{b}_{\text{error},i}^{\text{PA}} = \mathbf{b}_{\text{hist},i}^{\text{PA}} - \mathbf{b}_{\text{fore,hist},i}^{\text{PA}}$ to $\mathcal{D}_{\text{error}}$;
- 4: **end for**

Scheduling Stage*Aggregator:*

- 1: Infer $\mathbf{b}_{\text{fore}}^{\text{PA}}$ based on forecasts of uncertain DERs in the horizon to be scheduled;
- 2: Get $\mathcal{D}$ as Eq. (4.39) and submit;

Upper-level System:

- 3: Solve scheduling problems like Eq. (4.45) with the supporting algorithm in Algorithm 2; the solution of the aggregated power will meet the desired $1 - \epsilon$ reliability level.

Complexity analysis

Similar to the complexity analysis in Section 4.3.3, the only computationally demanding step is the derivation of $\mathbf{b}_{\text{fore}}^{\text{PA}}$ (Step 1) based on forecasts of uncertain DERs, by using existing methods that calculate the inner polytope approximation of the AFAPR in a deterministic setting (see Section 4.2.1). Notably, the proposed method (Steps 2 and 3 in the Scheduling Stage) has manageable computing complexity, which does not increase with the number of DERs nor with the size of the network, because the surrogate-polytope JCC is defined only for the aggregated power $\mathbf{p}^{\text{Agg}}$. Furthermore, as has been discussed in Section 4.3.3, the only computationally demanding part (Step 1) can be bypassed by generating $\mathcal{D}$ directly through a data-driven probabilistic forecasting framework.

4.6 Part II: case study

This section presents the case studies to validate the proposed method in Part II.

4.6.1 Case study settings

The case studies are carried out on the same IEEE-123 distribution feeder as Part I: case study. The `approxcdf` package [333] and the `stats.mvn.mvnun` function in SciPy are used to evaluate the CDF of multivariate Gaussian distributions. There are 20 PVs, 20 WTs, 10 ESs, 30 CLs, and 65 NLs. This section assumes that CL can be curtailed by 40%, which is the upper bound of the forecasted demand curtailment potential in the UK by 2050 [9]. The voltage limits are set to be 1 ± 0.05 p.u. Line current limits are not considered as the voltage limits can be the main issue for certain distribution networks [314]. The peak of the base demand in the whole network is around 3500 kW. To fully validate the proposed framework,

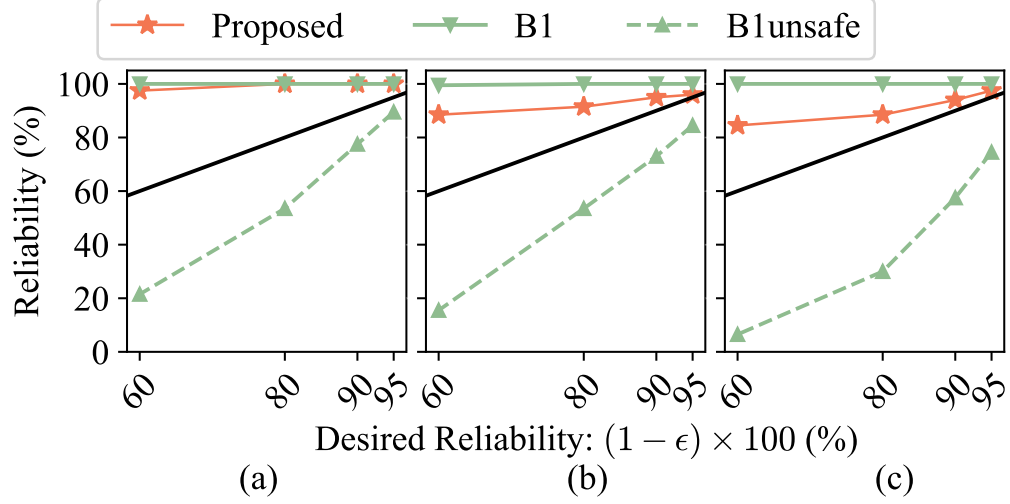


Figure 4.16: Out-of-sample reliability evaluation.

(a) 'LowRES', (b) 'HighPV', and (c) 'HighWT'. The black lines indicate the desired reliability levels.

this section creates three case studies: 1) 'LowRES': PV and WT capacities are both around 800 kW in total; 2) 'HighWT': the total WT capacity is set to 3500 kW; and 3) 'HighPV': the total PV capacity is set to 3500 kW. In the latter two cases, the distribution network is self-sufficient with local RG. The total ES capacity remains 3600 kWh across all case studies.

For the proposed surrogate-polytope method, this section sets the shape of the polytope to be the VB in (4.27) following Section 4.4.3. The datasets as well as the evaluation metrics used for this case study are the same as Part I: case study. Two benchmarks are set up for the case studies, which are 'B1' and 'B1unsafe' in Section 4.4.

4.6.2 Out-of-sample reliability results

Fig. 4.16 shows the out-of-sample solution reliability for (a) 'LowRES', (b) 'HighPV', and (c) 'HighWT', where the black diagonal lines are the desired out-of-sample reliability levels. It can be seen that both the proposed method and B1 can meet the reliability requirements for all the case studies, but B1unsafe cannot. This result confirms that ICC (B1unsafe) under the same risk level can lead to unreliable solutions, even though its static structure has led to conservativeness as explained in Section 4.4.2.

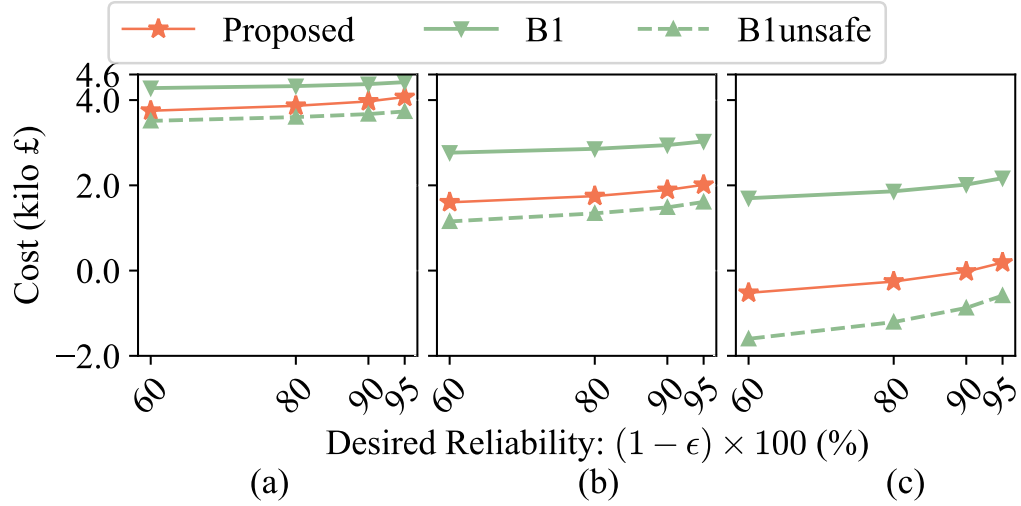


Figure 4.17: Out-of-sample cost evaluation.

(a) 'LowRES', (b) 'HighPV', and (c) 'HighWT'.

4.6.3 Out-of-sample cost results

Fig. 4.17 presents the out-of-sample cost for the three RES case studies. While B1unsafe achieves the lowest cost for all cases, its unreliability, as demonstrated in Section 4.6.2, compromises its suitability. The proposed method, being the second lowest in terms of cost, closely approaches the lowest costs achieved by B1unsafe. When considering the reliability evaluation detailed in Section 4.6.2, it can be concluded that the proposed method not only offers reliable solutions but also exhibits less conservativeness.

4.6.4 Scalability

To verify the applicability of the proposed method in wide-range problem scales, this section evaluates the out-of-sample cost and reliability under the total DER number (excluding NL) being from 50 to 120. Fig. 4.18 displays the results for the 'HighWT' case study. It can be seen that the conclusions from previous sections still hold: the proposed method can constantly meet the desired reliability requirements and is less conservative than B1. B1unsafe has a lower cost but cannot meet the reliability requirements for all cases.

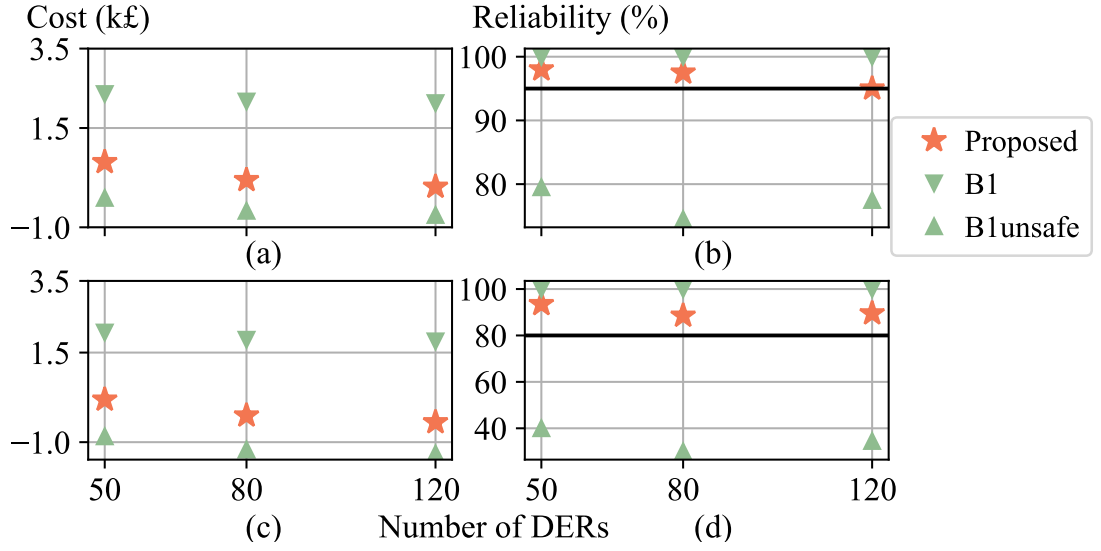


Figure 4.18: Out-of-sample evaluation for cost and reliability with different number of DERs.

First column: out-of-sample cost under different numbers of DERs with desired reliability levels (a) $1 - \epsilon = 0.95$ and (c) $1 - \epsilon = 0.8$; Second column: out-of-sample reliability with desired reliability levels (b) $1 - \epsilon = 0.95$ and (d) $1 - \epsilon = 0.8$, where the black horizontal lines indicate the desired reliability levels.

4.6.5 Overall computing time

As outlined in Section 4.5.2, the online calculations of the proposed method involve four steps: 1) computing $\tilde{b}_{\text{fore}}^{\text{PA}}$ using the bound shrinking algorithm in [95], which involves several mixed-integer linear programmes (MILP) solvable in 100-200 seconds in total for DER numbers from 50 to 120 with an AMD Ryzen9 5900X CPU. The case studies have relaxed the stopping bounds, capping the MILP solution time at 40 seconds; 2) employing the piece-wise linear approximation method with K^{PW} pieces from [94] to determine the aggregated DER cost function through $T \times K^{\text{PW}}$ parallelizable linear programs (4.3.2) taking 1.5-3.5 seconds each, leading to a total of 200-500 seconds without parallelisation; 3) fitting a multivariate Gaussian distribution for $\tilde{b}^{\text{PA}}$ with dimensions $P=96$, trivial due to the small scale of the VB polytope ($T=24$); and 4) applying the supporting hyperplane method in Algorithm 2. Fig. 4.19 illustrates the number of iterations and the computing time in applying the supporting hyperplane method to solve the JCC. It can be observed that both the number of iterations and the computing time are independent of the number of DERs. A lower level of reliability (a greater ϵ) requires more iterations due to the larger feasible region defined by the JCC. The computing time is around 20 seconds for a 95% reliability ($\epsilon = 0.05$), and around 40 seconds for an 80% reliability.

Implementing all the calculations sequentially results in a worst-case total computation time of approximately 12 minutes for 120 DERs in the IEEE-123 network with $T = 24$, manageable for a DA scheduling problem. Note that parallelisation could be explored to further speed up the calculation.

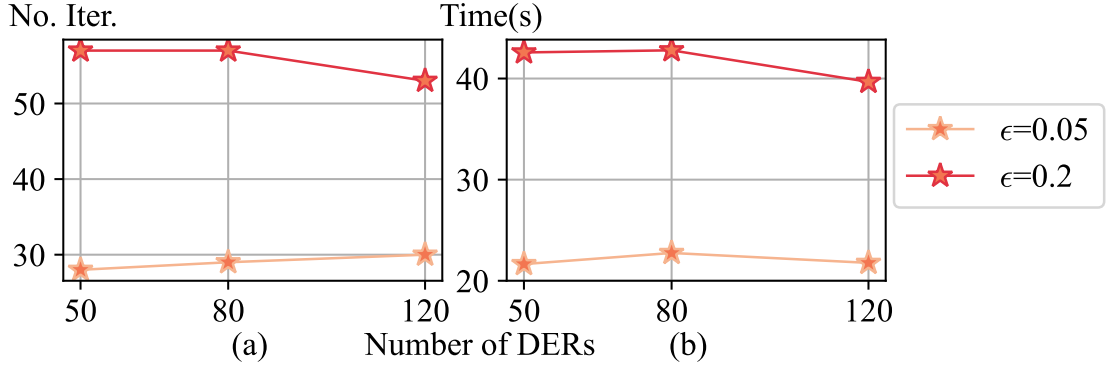


Figure 4.19: Computing efficiency evaluation for the hyperplane algorithm with different number of DERs.

(a) The number of iterations (No. Iter.) in applying the proposed supporting hyperplane algorithm to solve the proposed surrogate JCC under different numbers of DERs; (b) The computing time in applying the proposed supporting hyperplane algorithm to solve the proposed surrogate JCC.

4.7 Discussion

This thesis provides further discussions on the proposed method and existing approaches for handling uncertainty in aggregation.

4.7.1 Comparison with the two-stage RO framework

The proposed polytope method enables the efficient integration of WDRJCC or classical JCC with a two-stage structure, resulting in an inner approximation of the jointly reliable AFAPR defined in Eq. (4.24). This section demonstrates that the jointly reliable AFAPR is closely related to existing research that uses RO for reliable aggregation. For convenience, the jointly reliable AFAPR studied in this chapter is restated below:

$$\Omega^{\text{Agg}}(\epsilon) = \left\{ p^{\text{Agg}} \mid \mathbb{P} \left(p^{\text{Agg}} \in \Omega^{\text{Agg}} \right) \geq 1 - \epsilon \right\} \quad (4.54a)$$

$$= \left\{ p^{\text{Agg}} \mid \mathbb{P} \left(\exists x^{\text{DER}}, \text{ Eqs. (4.1)–(4.22)} \forall \psi, i_d, t \right) \geq 1 - \epsilon \right\}. \quad (4.54b)$$

This definition is conceptually similar to existing AFPR work that employed two-stage RO (or adaptive RO) [161–164], as reviewed in Section 2.2.3. The two-stage RO framework ensures that for any desired aggregated power, and for any realisation of uncertainty, there always exists a disaggregatable DER action x^{DER} that satisfies all DER constraints and network constraints.

As evident from Eq. (4.54b), the only difference is that the jointly reliable AFAPR here does not guarantee satisfaction for all possible realisations of uncertainty. Instead, it guarantees satisfaction with a probability of at least $1 - \epsilon$, making it less conservative than the two-stage RO framework.

4.7.2 Limitation

A limitation exists in the disaggregation process for both the proposed method and the two-stage RO approach when $T \geq 2$. As shown in Eq. (4.54b), while the proposed method guarantees the existence of the second-stage DER actions x^{DER} , the derivation of x^{DER} remains challenging as the two-stage framework does not perfectly capture the real-world multi-stage implementation process. In the case studies presented, the disaggregation x^{DER} is performed after knowing the DER realisation for all T time steps, as per Eq. (4.37). However, this approach is not applicable in real-world applications because it violates the non-anticipativity constraint in the real-world multi-stage dispatch process; in other words, the decision-maker should be unsure about the realisation in future time steps, so decisions cannot be performed based on these future information [290]. It is noteworthy that this issue does not arise for $T = 1$, where the multi-stage dispatch reduces to a two-stage problem as modelled in the proposed method. Common examples include the single time-step P-Q chart, as studied in [24, 161–163], which is useful when flexibility is only exploited for a single time-step, or when there are no time-coupling constraints, e.g., ES energy bounds, such that a T -time-step aggregation simplifies to T single time-step aggregations.

To the best of the author's knowledge, no existing research adequately addressed this limitation. Among the existing studies that derived reliable AFPRs using CCs, the methods can be categorised as follows:

1. Method that uses a static JCC framework, as in [168], where the disaggregation x^{DER} is fixed throughout all future realisation of uncertainties. This is the B1 and B1unsafe benchmarks in this chapter. As has been discussed in Section 4.4.2, this static framework leads to over-conservativeness and there will be ill-posed equality constraints that cannot be properly handled.
2. Method that imposes CCs on the maximum or minimum capability of the aggregated power [94, 95]. This method is the B2 and B2unsafe benchmarks in this Chapter. As has been discussed in Section 4.4.2, this method still shares the two-stage framework; consequently, the same $T \geq 2$ limitation persists. Furthermore, as demonstrated in this chapter, this method is either more conservative (B2) or less reliable (B2unsafe) compared to the proposed method.
3. Method that transformed the ill-posed equality constraint (discussed in point 1 here) into inequality constraints by exploiting vector transformations via basis vectors [170]. However, this transformation also causes the DER power x^{DER} to disappear, and x^{DER} needs to be recovered through an optimisation problem based on a known realisation of uncertain DERs (The original paper [170] used the sample mean value). Consequently, this method is still within the two-stage framework and the same limitation persists.

4.7.3 Potential solution

Addressing the identified limitation necessitates the integration of the real-world multi-stage structure into DER aggregation, incorporating non-anticipativity constraints [290]. Similar to the approach discussed in Remark 7, one may first formulate a JCC with a non-anticipative multi-stage structure that includes all DER constraints and network constraints, and subsequently apply aggregation algorithms to calculate the reliable AFPR. However, as also noted in Remark 7, the convexity or linearity requirements, coupled with the additional computational complexity imposed by common aggregation algorithms, present a significant computational challenge. Future work may focus on developing more tractable multi-stage JCC reformulations and/or enhancing DER aggregation algorithms to address these challenges effectively.

4.8 Concluding remarks

This chapter presented a novel surrogate polytope method for modelling the jointly reliable AFAPR over the entire scheduling horizon, accounting for linear network constraints and DER constraints, including time-coupling effects. The proposed method begins with a polytope inner approximation with a fixed slope in a deterministic setting. By extending this to the probabilistic case, the polytope approximation converts the original complicated large-scale two-stage JCC into a single-stage RHS-JCC, leading to an analytical reformulation for WDRJCC and a solution scheme for classical JCC, both of which are efficient and exact. The computational complexity of the integration of WDRJCC and JCC is low, which does not increase with the number of DERs nor with the size of the network.

It has been demonstrated that the proposed method provides an inner approximation to the original intractable jointly reliable AFAPR. Furthermore, through comparisons with benchmarks across various case studies, the method has been shown to yield a less conservative yet still reliable AFAPR. The satisfaction of network constraints was validated using the nonlinear Z-Bus power flow simulation, reinforcing the reliability of the proposed method.

Up to this point, the thesis has addressed the first research question related to the technical challenge of unlocking DER flexibility, namely *Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering the DER uncertainty?* This is achieved by proposing SFLA for speeding up general power system DER scheduling problems under uncertainty (the previous chapter) and a novel method to deal with uncertainty in DER aggregation (this chapter). In addition to the technical challenge, as discussed in Section 1.1.3, understanding the impact of DERs also represents a critical dimension in unlocking the full value of DER flexibility. However, two major types of DERs, namely HPs and data centres, remain under-explored, which are the focus of the subsequent chapters.

Heat Pump Impact on Fuel Poverty and Distribution Networks

As discussed in Chapter 1, the technical, economic and social impacts of DER also represent a crucial dimension that supports the design of effective and efficient roadmaps to realise DER flexibility. However, HPs and data centres remain under-investigated. Given the substantial power demand that HPs and data centres are projected to contribute to future power systems (as illustrated in Fig. 2.2), this knowledge gap poses a significant challenge in unlocking the full value of DER flexibility. Specifically, this chapter examines HPs and addresses the second research question:

Q2: What are the fuel poverty and distribution network impacts of replacing domestic off-gas heating systems with HPs?

As discussed in Sections 1.1.3 and 2.3.4, HPs represent a key technology for heating electrification. With an efficiency of around 300%, HPs offer significant potential to reduce fuel poverty—a state where households struggle to afford the energy required to heat their homes comfortably. Fuel poverty has been a critical social issue for decades, and the 2022–2023 energy crisis, triggered by the COVID-19 pandemic and the Russia–Ukraine war, has further exacerbated this issue, raising social awareness and concern around it. The high efficiency of HPs makes them a promising solution to the issue of fuel poverty, particularly for off-gas homes that rely on costly off-gas heating options, such as ERH, heating oil, or LPG. However, as noted in Section 2.4.2, there is a lack of research on the fuel poverty impact and network impact when HPs are to replace these expensive off-gas heating options.

To address this research gap, this chapter proposes a method that supports a detailed analysis for the potential benefits of replacing off-gas heating with HPs in terms of reducing fuel poverty and GHG emissions, while also evaluating the associated electricity distribution network upgrade cost at a regional level. Using processed national household datasets from England and Scotland, alongside electricity substation information from UK DNOs (DNOs operate and manage the electricity distribution system that delivers power to most end-users), a detailed examination is conducted for the 9 government regions in England and the 32

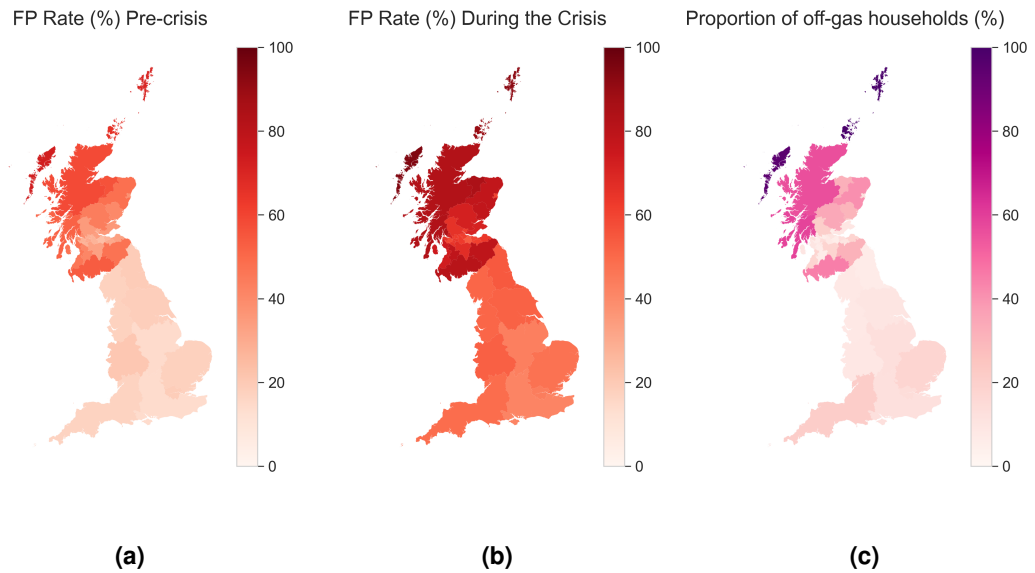


Figure 5.1: Fuel poverty estimations and proportions of off-gas households.

(a) Fuel poverty (FP) rate (%) estimation before the energy crisis. (b) FP rate (%) estimation during the energy crisis. (c) The proportion of households using off-gas heating options.

local authorities (LoAs) in Scotland. It is demonstrated that HPs replacing off-gas heating can reduce fuel poverty for both normal period and energy-crisis period, and certain regions can suffer high network upgrade cost. The analysis demonstrates the positive social benefits of HP rollout and also the urgent need for HP flexibility to alleviate network stress.

The findings presented in this chapter offer valuable insights for policymakers aiming to address fuel poverty and achieve decarbonisation targets. Moreover, the methodology developed can be adapted for application in other countries given appropriate datasets.

5.1 Methods

This section describes the proposed method to estimate both the fuel poverty impact and the network upgrade impact imposed by HPs. The proposed method leveraged a combination of datasets from various public sources and achieves the evaluation for the 9 government regions in England and the 32 LoAs in Scotland. To facilitate understanding, these data sources are summarised in Table B.1 in Appendix B. In addition, as a low-carbon heating option, HPs can also deliver carbon reduction benefits, which are also evaluated in this chapter; the corresponding evaluation method is presented in this section as well.

This section is organised into the following three parts:

1. Section 5.1.1 describes the method for estimating fuel poverty, along with the datasets supporting the corresponding case studies.

2. Section 5.1.2 outlines the approach for estimating GHG emissions.
3. Section 5.1.3 details the method for estimating distribution network upgrade costs, as well as the datasets used to support the associated case studies.

5.1.1 Estimation of fuel poverty

This section details the proposed process for estimating fuel poverty. The key information extracted from the relevant datasets is introduced first, followed by a presentation of the official definition of fuel poverty and the corresponding estimation method.

Household dataset

Two national surveys, the English Housing Survey (EHS) and the Scottish House Condition Survey (SHCS), provide processed household interview data in [44, 334]. The SHCS dataset [334] spans from 2012 to 2019 (inclusive) and includes approximately 3,000 samples per year. Each sample represents an anonymous household and contains information on:

- Household adjusted net income: the income of all adults in the household after deduction of the income tax payable on that income, the national insurance contributions, and housing costs [40].
- Modelled energy consumption: the required annual energy in kWh for the household, including energy for heating, cooking, and lighting.
- Primary heating fuel type: the primary heating fuel for this household. Examples include mains gas, electricity, heating oil, LPG, etc.
- LoA: the LoA where this anonymous household locates.
- Sampling weights: the weight is applied to each household interview sample, so the weighted statistics of the whole dataset can be representative of the whole nation.

This information is sufficient for subsequent fuel poverty estimation. Data from 2017 to 2019 are used to obtain a more accurate estimation for each LoA, as recommended in their methodology report; otherwise, the lowest number of data samples for a single LoA would be only 54.

The English dataset follows a similar format, and this chapter uses the most recent data from 2019, which contains approximately 12,000 samples. The minimum number of samples per region is 729, so there is no need to include multiple years as in the case of Scotland. Unlike the Scottish data, which links each sample to a specific LoA, each sample in the English dataset only indicates the corresponding government region, out of nine total (the regions, formerly known as the government office regions, are the highest tier of sub-national division in England, which includes South East, London, North West, East of England, West Midlands, South West, Yorkshire and the Humber, East Midlands, and North East). This classification (which is less granular than LoAs) is the lowest geographical level for the England fuel poverty analysis results in this thesis.

Fuel poverty definition

This chapter uses the “10% definition” of fuel poverty: A household is considered to be in fuel poverty if its necessary annual fuel cost is more than 10% of the household adjusted annual net income. This applied 10% definition is popular in the UK, and is being used (or was used) as official measurements for several regions with nuance differences. For example, Wales is using the 10% definition and considers a household in fuel poverty if its necessary fuel cost is more than 10% of the household's **total income** [335]. Scotland is using the 10% definition as well, with two minor differences on 1) the household's income is defined as the household **adjusted net income** and 2) a fuel-poor household should also satisfy that the household's remaining adjusted net income is insufficient to maintain an acceptable standard of living [40]. England now uses the “Low Income Low Energy Efficiency” as the fuel poverty criterion [54], in which a household is considered fuel-poor if it has both low income and low energy efficiency (a Fuel Poverty Energy Efficiency Rating (FPEER) of band D or below [336]); however, the 10% definition was used before 2012. Note that, according to [40], “adjusted” means adjusted by deducting housing costs; “net income” means the income of all adults in the household after deduction of (i) the income tax payable on that income, and (ii) the national insurance contributions for which those adults are liable under Part 1 of the Social Security Contributions and Benefits Act 1992.

The “10% definition” used in this thesis is closest to the Wales definition. The use of the **adjusted net income** instead of the **total income** in the Wales definition is due to data availability. However, even using a different income definition, the “10% definition” used in this thesis still reflects the affordability of household heating, which aligns with the fundamental issue that the fuel poverty concept wants to point out. The proposed fuel poverty estimation framework can also flexibly switch to other definitions with necessary datasets.

Estimating fuel poverty

With the “10%” fuel poverty definition, the key to fuel poverty estimation is the fuel cost estimation, after which the fuel poverty status for each household can be calculated by comparing to the household income provided by the dataset.

The annual fuel cost C_{AF} for each household can be calculated as:

$$C_{AF} = \hat{C}_{\text{heat}} \times E_{\text{SW}} + \hat{C}_{\text{elec}} \times E_{\text{other}}, \quad (5.1)$$

where E_{SW} refers to the energy consumption in kWh-heat for space and water heating with a corresponding unit heating cost $\hat{C}_{\text{heat}}$, and E_{other} refers to other energy consumption including cooking and lighting with a unit cost $\hat{C}_{\text{elec}}$.

From Eq. 5.1, it can be seen that fuel prices play an important role in calculating the fuel cost. Based on [43] and [44], the “other energy consumption” is mainly contributed by electricity, and thus this thesis sets the corresponding unit cost $\hat{C}_{\text{elec}}$ as the standard electricity tariff [337]. Then it remains to collect information for the unit heating cost $\hat{C}_{\text{heat}}$ that corresponds to different primary heating option.

It is more common to find data on the unit fuel price $\hat{C}_{\text{fuel}}$, rather than the desired unit heating cost $\hat{C}_{\text{heat}}$ in the UK. However, these two pieces of information can be mutually converted by knowing the heater efficiency η_{heater} and through the following relationship:

$$\hat{C}_{\text{heat}} = \hat{C}_{\text{fuel}} / \eta_{\text{heater}}. \quad (5.2)$$

For the unit fuel price $\hat{C}_{\text{fuel}}$, common heating fuels include electricity, mains gas, heating oil, coal, LPG, and biofuels in the UK [43, 44]. Electricity and mains gas prices are sourced from BEIS quarterly energy prices [337], while heating oil and coal prices are obtained from the BEIS Consumer Price Index [338]. Both datasets represent national statistics. LPG and biofuel prices are derived from Nottenergy [339], which collects price information from suppliers within the East Midlands region of England. Note that, England’s fuel poverty data specifies three categories of heating fuels: 1) mains gas, 2) electricity, and 3) other fuel types; thus extra steps are required to estimate the unit heating cost for these three heating fuels. As the fuel poverty dataset does not specify whether the electric heating is EDH or ESH, this chapter assumes the price for ‘electricity’ option as the weighted average price of ESH (5/11) and EDH (6/11), based on their share in UK household electric heating surveyed by BEIS [56]. ESH benefits from the UK Economic-7 tariff, and this chapter follows [53] and assumes that 90% of the daily consumed energy is charged over 0:00 and 7:00 am, the low-price period in Economic-7. This chapter defines the fuel price for the “other fuel types” option in England as a weighted average of all non-gas fuels (excluding electricity), with weights determined by the household heating proportions in England based on 2021 England census data [340].

For the heater efficiency η_{heater} , this chapter sets the heater efficiency at 90% for gas, LPG, and oil boilers, 75% for coal boilers, 100% for electric heating, and 300% for HP, as suggested in [53, 339]. The unit heating cost $\hat{C}_{\text{heat}}$ is finally calculated through Eq. (5.2).

As for the energy consumption E_{SW} and E_{other} , the SHCS dataset only provides the total energy consumption $E_{\text{tot}} = E_{\text{SW}} + E_{\text{other}}$. However, based on the official SHCS summary report [43], on average 87% of energy is used for heating in Scottish households. Therefore the two energy consumption quantities are calculated as $E_{\text{SW}} = 0.87E_{\text{tot}}$ and $E_{\text{other}} = 0.13E_{\text{tot}}$ for Scottish households. The EHS dataset only provides the modelled energy cost C_{SW} and C_{other} in £ rather than the desired energy consumption in kWh. This chapter estimates the corresponding energy consumption in kWh using the unit heating cost in the same year (2019) of creating the England dataset as $E_{\text{SW}} = C_{\text{SW}} / \hat{C}_{\text{heat}}^{2019}$ and $E_{\text{other}} = C_{\text{other}} / \hat{C}_{\text{elec}}^{2019}$.

Note that, when calculating the annual fuel cost C_{AF} in (5.1), this chapter does not consider fixed costs, such as standing charges for metered fuels or maintenance and delivery costs for non-metered fuels, because the primary focus of this thesis is on energy usage costs. Moreover, this chapter centres on HP replacements for off-gas heating, which means that considering fixed costs would essentially make HP a more economical option, as off-gas non-electric heating fixed costs would be reduced (almost all properties use electricity, so they have already been paying the electricity standing charge). The estimated annual maintenance cost is £90-£125 for gas boilers, £100-£180 for oil boilers, and £80-£125 for LPG boilers [341]. The maintenance cost for HP can be around £175, which is slightly more expensive, but the gap can be wiped out by not paying the fixed cost for off-gas non-electric after HP replacement [53]. Government benefits like the Warm Home Discount (WHD) are also not included for the same reason.

Although, by definition, fuel poverty is determined solely based on the running cost of the primary heating option, the high upfront cost of HPs—typically three to four times higher than that of gas boilers—raises concerns that the calculated fuel poverty benefit of HPs may obscure the financial downside associated with their high capital cost. Therefore, this section proposes a method to account for the upfront cost of HPs when estimating fuel poverty, by introducing the concept of a corrected annual fuel cost, C'_{AF} , defined as:

$$C'_{AF} = C_{AF} + C_{AmoF}, \quad (5.3)$$

where C_{AmoF} refers to the first-year payment towards the amortised HP upfront cost. The first-year payment under varying interest rates can be obtained from an online calculator [342] or calculated using the formula provided in [343]. The intuition behind Eq. (5.3) is to assume that the HP is purchased through an amortised payment plan. In fact, Appendix B.3 demonstrates that Eq. (5.3) is equivalent to the calculation using the levelised lifetime cost method, which is widely employed in measuring per-unit energy costs that account for capital investments [344].

Subsequent case studies will evaluate the HP fuel poverty impact under both the original fuel cost C_{AF} and the corrected fuel cost C'_{AF} .

5.1.2 Estimation of greenhouse gas emissions

The domestic heating of a household generates an amount of GHG emissions E_{GHG} , which can be calculated by:

$$E_{GHG} = E_{SW}/\eta_{heater} \times C_{GHG}, \quad (5.4)$$

where the GHG Emissions per kWh energy (GHG factor C_{GHG}) by combustion comes from the UK government report [345]. Note that this chapter focuses on the GHG emissions caused by domestic heating only, excluding other household activities that generate GHG emissions.

5.1.3 Estimation of distribution network upgrade costs

The widespread adoption of electric HPs could trigger upgrades for distribution network infrastructures, which can be a financial burden for DNOs and eventually result in higher customer bills. This chapter starts with the cost estimation for all primary substations (PS) in Section Estimating the PS upgrade cost. PS is a crucial part of the UK distribution network and is also at the lowest voltage level with publicly available data for most UK DNOs. This chapter then scales the upgrade costs of PSs to the whole distribution network as detailed in Section Scaling up network upgrade cost.

Substation dataset

DNOs in the UK publish PS information on their corresponding websites listed in Table B.1 in Appendix B. The required information includes 1) existing peak demand (net of local generation contribution), 2) firm capacity in MVA, 3) power factor (PF) (informally speaking, PF measures how much electrical power can be actually consumed by end-users), 4) longitude and latitude which are used to infer their corresponding geographical regions. These information may be contained in single or multiple files for a DNO. All the information are extracted and merged to get the processed PS dataset. Entries with missing information except for PF are dropped, resulting in 4210 out of 4347 PSs, an approximately 3.15% reduction. Where PF is missing, it is set to 0.9855, which is the average values of PSs with PF. As this chapter specifically focuses on England and Scotland, PSs in these two countries are extracted, and there are 3,891 PSs in total.

Estimating the PS upgrade cost

Figure 5.2 summarises the proposed method for estimating the PS-level network upgrade cost of HPs replacing off-gas domestic heating. For the i -th PS, its upgrade cost $C_{PS,i}$ to accommodate the new demand of HPs is calculated as:

$$C_{PS,i} = S_{E,i} \times \hat{C}_{PS}, \quad (5.5)$$

where $S_{E,i}$ is the exceeding capacity in MVA (the amount of electrical power that a PS cannot supply). $\hat{C}_{PS}$ is the unit cost of each MVA upgrade of a PS derived in Step (vii). The exceeding capacity $S_{E,i}$ of a PS can be further calculated as:

$$S_{E,i} = \overline{D}_{new,i} - S_{F,i}, \quad (5.6)$$

where $\overline{D}_{new,i}$ is the new peak demand after the HP rollout (Step (v)) for this PS, and $S_{F,i}$ is the firm capacity of the PS described in Step (vi).

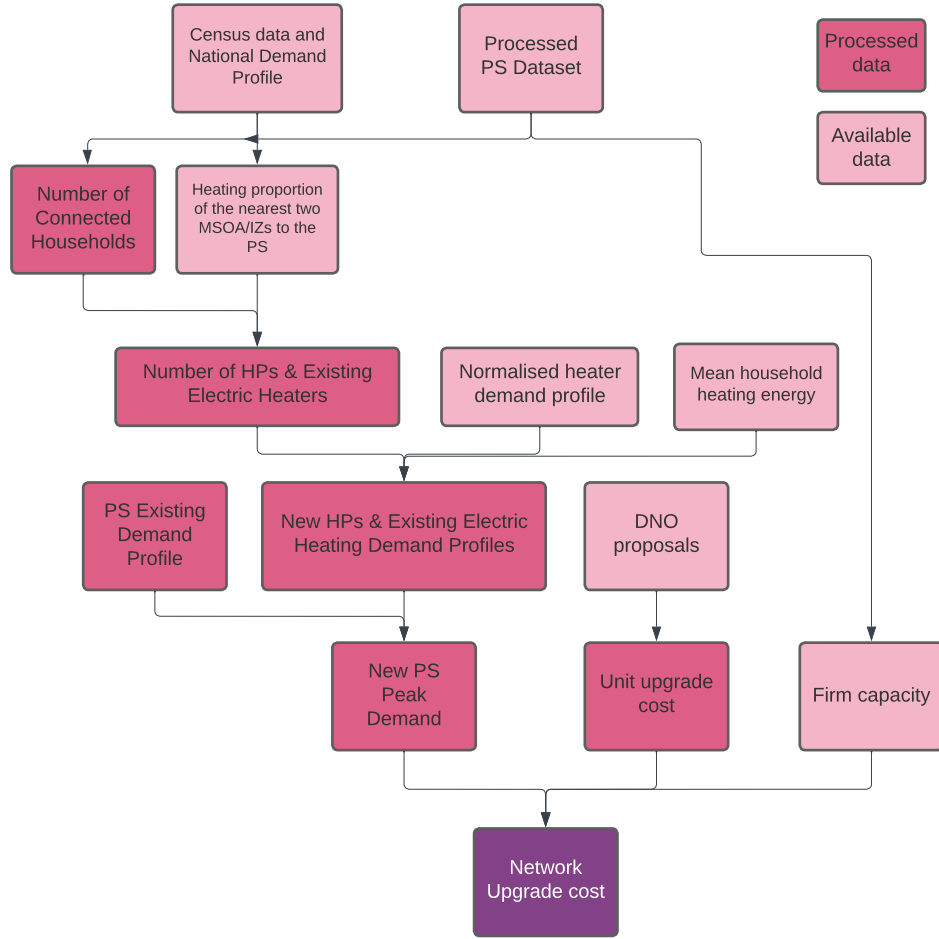


Figure 5.2: Flowchart for estimating the PS upgrade cost.

The following content describes how the new PS peak demand is derived based on the bottom-up order, as shown in the flowchart. The description focuses on the i -th PS of a region (the minimum region unit in the fuel poverty estimate of this thesis), and this chapter assumes this region has a total N_{PS} PSs. One can easily repeat the calculation for each PS and make a summation to get the regional PS upgrade cost.

(i). Number of connected households

The first step is to calculate the number of connected households $N_{house,i}$ for the i -th PS under evaluation. The exact number of $N_{house,i}$ is not provided for most of the UK DNOs, and here $N_{house,i}$ is estimated by dividing the total number of households N_{house} in this region by the PSs' existing peak demand $\bar{D}_i$ from the PS dataset, which implicitly neglects the effect of other demand than the domestic sector, e.g., industrial sector. In other words,

$$N_{house,i} = N_{house} \times \frac{\bar{D}_i}{\sum_i^{N_{PS}} \bar{D}_i}. \quad (5.7)$$

As a toy example, suppose region A with 10,000 households has two PSs: PSa and PSb. The peak demand in PSa is 4 MW, while that in PSb is 6 MW. Then based on the proposed method, there will be 4,000 households connected to PSa and 6,000 to PSb.

(ii). Number of HPs and existing electric heaters

This step calculates the number of required HPs to replace all the off-gas heaters $N_{HP,i}$ that equals to the number of off-gas heaters, and the number of existing electric heaters $N_{EH,i}$ connected to the i -th PS.

An ideal approach would be to first identify the supply area of each PS and then combine this information with census data on household heating proportions. However, most DNOs do not disclose the exact supply regions in a processable format. Based on the estimation in Step (i), the mean and median number of connected households for PSs in England are 7,445 and 6,320 respectively. For Scotland PSs, the mean and median are 3,185 and 2,864 respectively. In census data, the geo-levels that are closest to the number of PS-level households are Middle Layer Super Output Area (MSOA) for England and Intermediate Zone (IZ) for Scotland. The mean and median number of households for an MSOA are 3,418 and 3,308. The mean and median household numbers of an IZ are 1,855 and 1,802. Therefore, a PS can roughly support two MSOAs or two IZs in England or Scotland.

This observation suggests that the household heating component for PS i can be derived by combining the household heating component data for the two nearest MSOAs in England and the two nearest IZs in Scotland (available in Census [340, 346]). For $N_{HP,i}$, by knowing the number of off-gas heaters in the two nearest MSOAs/IZs denoted as $N_{HP,i,1}$ and $N_{HP,i,2}$, it follows:

$$N_{HP,i} = N_{house,i} \times \frac{N_{HP,i,1} + N_{HP,i,2}}{N_{house,i,1} + N_{house,i,2}}, \quad (5.8)$$

where $N_{house,i}$ stands for the total number of connected households of the i -th PS estimated in Step (i). $N_{house,i,1}$ and $N_{house,i,2}$ are the total number of households in the two nearest MSOAs/IZs from the census dataset. The number of existing electric heaters $N_{EH,i}$ can be calculated in the same manner.

For example, consider a PS estimated to supply 10,000 households. The two closest IZs are IZ1 and IZ2. IZ1 has a total of 5,000 households, with 1,000 using off-gas heating and 300 using electric heating. IZ2 also has 5,000 households, 2,000 of which use off-gas heating and 700 using electric heating. In this case, the number of HPs for PS is calculated as $10,000 \times (1,000 + 2,000) / (5,000 + 5,000) = 3,000$, and the number of existing electric heaters is $10,000 \times (300 + 700) / (5,000 + 5,000) = 1,000$.

The fuel poverty datasets also have information on the number of households in each heating option, but for the region (LoA for Scotland and Government Regions for England) in total. For better consistency, this chapter subsequently scales the estimated number of various heaters for each PS so that the summed number of heaters of all the PSs for this region matches the fuel poverty dataset.

(iii). New HPs and existing electric heating demand profiles

An open dataset in [347] provides a typical normalised HP and EDH electricity demand profile, with a one-year duration and half-hour resolution. Take the HP demand profile as an example, this normalised demand profile can be denoted as a vector $\hat{\mathbf{D}}_{\text{HP}} = [\hat{D}_{\text{HP},1}, \dots, \hat{D}_{\text{HP},17520}]$. Here “normalised” means the one-year demand profile sum up to 1 kWh of electricity usage, i.e., $\sum_t \hat{D}_{\text{HP},t} = 1$. This demand profile is derived based on the aggregated load profile; thus it incorporates the After Diversity Maximum Demand (ADMD), which is the average peak load per customer after accounting for the fact that not all customers reach their maximum demand at the same time. This demand profile allows for a more accurate peak demand estimate when multiple HPs are connected to a PS.

The heating boilers may be turned off when residents are away from their homes. The two national fuel poverty datasets (see Section 5.1.1) take this behaviour into account when modelling the household heating energy, resulting in two categories of heating patterns:

- 9 hours on working days (6-8:00 am and 17-24:00) and 16 hours on weekends (8-24:00), which is the heating pattern in the Standard Assessment Procedure (SAP) [43].
- 16 hours (8-24:00) every day, considering that a household member stays at home daily to care for children or has vulnerable members.

The two national fuel poverty datasets also assumed a heating season spanning from October to May (heaters are operational only within this period). Under the heating pattern assumptions above, the national fuel poverty datasets estimate the annual heating energy consumption per household, and this annual consumption is then used for scaling the normalised heating demand profiles in this section. However, [347] did not consider these heating pattern when creating the normalised heating demand profiles, and thus these profiles can correspond to different annual heating energy consumption. To mitigate this inconsistency, the half-hourly heating demand profiles are cut and reshaped to align with the heating pattern assumptions in the fuel poverty datasets.

Note that, as the exact household composition connected to each PS is unavailable, this chapter assumes that all households follow the first heating regime, i.e., 9 hours on weekdays and 16 hours on weekends. Given the same amount of heating energy, a reduction in the heating period will lead to an increase in heating demand, leading to a conservative estimate of network upgrade costs.

The reshaped annual heating demand profiles of individual HPs are then aggregated to the PS level. For the i -th PS, this chapter calculates the total new HP heating profile $D_{HP,i}$ by

$$D_{HP,i} = N_{HP,i} \cdot \frac{E_{\text{heat}}}{\mu_{HP}} \cdot \hat{D}_{HP}, \quad (5.9)$$

where $N_{HP,i}$ is the number of required HPs to replace all the off-gas heaters in Step (ii). E_{heat} is the mean household heating energy for the region, as can be determined from the fuel poverty dataset (Section 5.1.1). μ_{HP} is the efficiency, which is the seasonal performance factor (SPF) for HPs here (The SPF is a more appropriate efficiency metric in this context than the coefficient of performance (COP), as SPF accounts for the seasonal variation in COP). This chapter assumes a PF of 1 for HPs as suggested by [348], which means that HPs do not contribute reactive power.

It remains to calculate the existing electric heating demand. Following the settings in the fuel poverty analysis in Section 5.1.1, the existing electric heating demand $D_{EH,i}$ is calculated as $6/11 \times \text{EDH demand} + 5/11 \times \text{ESH demand}$, and the ESH demand is derived by reassigning the EDH demand so that the energy consumed between 0:00 and 7:00 am accounts for 90% of the daily consumed energy. Following the same way as $D_{HP,i}$, this chapter calculates the total demand profile $D_{EH,i}$ starting from its normalised demand profile $\hat{D}_{EH,i}$.

(iv). PS existing demand profile

Due to data unavailability, the PS existing demand profile is estimated by scaling down the UK national electricity demand profile [349] to the PS peak demand. The UK demand profile in year 2019 is selected because this is the latest year with full-year data. The one-year existing electric active power demand profile $D_{P,i}$ for the PS is obtained after scaling. The reactive power demand profile $D_{Q,i}$ can be calculated as:

$$D_{Q,i} = D_{P,i} \cdot \sqrt{\frac{1 - \alpha_{PS,i}^2}{\alpha_{PS,i}^2}}, \quad (5.10)$$

where $\alpha_{PS,i}$ is the PF of the i -th PS recorded in the substation datasets.

(v). New PS peak demand

Given 1) the new HP demand profiles $D_{HP,i}$ due to replacing off-gas heaters in Step (iii), 2) the existing electric heating demand profile $D_{EH,i}$ in Step (iii), and 3) existing active and reactive demand profiles $D_{P,i}$ and $D_{Q,i}$ in Step (iv), the new yearly demand profile D_{new} (apparent power in MVA) of the i -th PS can be calculated as:

$$D_{\text{new}} = \sqrt{(D_{Q,i})^2 + (D_{P,i} + D_{HP,i} - D_{EH,i})_+^2}, \quad (5.11)$$

where $D_{P,i} + D_{HP,i} - D_{EH,i}$ represents the new active power demand profile, and the negative sign in $-D_{EH,i}$ is because the existing electric heating (ESH and EDH) is also replaced with HPs. The $(\cdot)_+$ function sets negative elements to zero. In the case study of this chapter, the new active power demand profiles can contain negative values for 0.155% of time-steps for all PS in total throughout the year. This could be due to a slight over-estimate of the number/demand of existing electric heaters for each PS.

The new peak demand $\overline{D}_{new,i}$ after the HP rollout is calculated as:

$$\overline{D}_{new,i} = \max_{D_{new,t} \in \mathbf{D}_{new}} D_{new,t}. \quad (5.12)$$

(vi). Firm capacity

Based on a UK DNO Northern Powergrid [350], firm capacity $S_{F,i}$ refers to the capacity of a substation that is immediately available following the occurrence of a first circuit outage. Power system operation has a high requirement on reliability and resilience, so firm capacity can well measure if a PS can supply electricity over peak time under extreme conditions. This information has been included in the processed PS dataset as described in Section 5.1.3.

(vii). Unit upgrade cost

Several PS upgrade proposals were obtained from a UK DNO, which provided the installed new capacity and the associated cost. By averaging these, this chapter determined the upgrade cost per exceeding MVA $\hat{C}_{PS}$ to be £236,900/MVA. It is important to note that this is only a high-level estimation which aims to provide insights and indications rather than a precise figure which would depend on local network details. A detailed analysis of PS upgrade cost is complex and requires a granular evaluation by the corresponding DNO.

Finally, based on new peak demand $\overline{D}_{new,i}$ after the HP rollout in Step (v), the firm capacity $S_{F,i}$ in Step (vi), and the unit upgrade cost $\hat{C}_{PS}$ in Step (vii), the exceeding capacity $S_{E,i}$ can be calculated through Eq. (5.6) and then the upgrade cost $C_{PS,i}$ of the i -th PS can be calculated through Eq. (5.5).

Scaling up network upgrade cost

The previous section has estimated the total PS upgrade cost. This chapter scales the PS upgrade cost up to the whole distribution network cost based on DNO summary statistics. Fig. 5.3 visualises the network cost components for all UK DNOs, based on which the cost scaling factors for each DNO can be inferred.

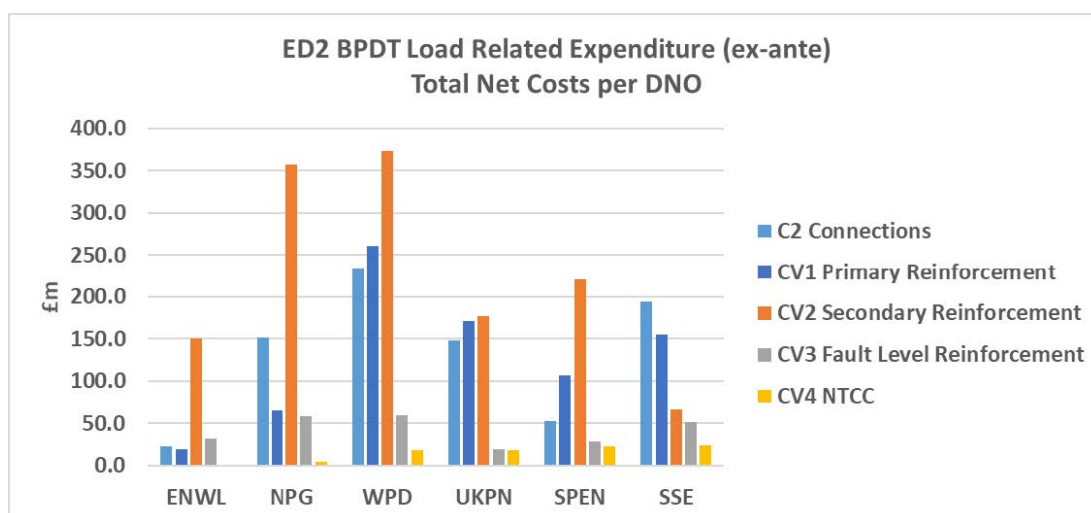


Figure 5.3: Breakdown of load related expenditure by UK DNOs.

The total Load Related Expenditure (LRE) is the total distribution network upgrade cost, and “CV1 Primary Reinforcement” is the PS network upgrade cost estimated in Section 5.1.3. Data is sourced from the RIIO-ED2 in [351]. The cost terms are defined in the RIIO-ED2 methodology document [351] and are explained below:

“C2 Connections” refers to the provision of new or upgraded network points of connection which can be metered or unmetered connections with the end customer. It includes reinforcement costs associated with the connections work.

“CV1 Primary Reinforcement” covers reinforcement activity undertaken to resolve capacity constraints on the primary network (33kV and above), which includes substation and circuit (cable).

“CV2 Secondary Reinforcement” is work carried out on the secondary network (HV and LV) to enable new load growth. This includes 1) transformer and circuit reinforcement, 2) proactive service reinforcement, and 3) flexibility services and other reinforcement activities.

“CV3 Fault Level Reinforcement” covers work carried out on the existing network where the primary objective is to alleviate fault level issues associated with switchgear or other equipment.

“New Transmission Capacity Charges (CV4 NTCC)” are the charges payable for projects initiated by the DNO for increased capacity at existing transmission connection points or for new transmission connection points, but carried out by transmission licensees.

5.1.4 Summary of assumptions

For the convenience of reading, this section summarises the key assumptions made in this chapter and discusses their potential impacts. Detailed reasons and justifications for these assumptions can be found in the corresponding sections of this chapter.

1. Section 5.1.1: The fuel poverty definition used in this chapter is the “10% definition” based on adjusted net income. Currently, England does not use this definition, Scotland applies an additional condition, and Wales uses total income. This choice is due to data availability. It is not anticipated that the change of definition would significantly affect the results of the HP benefits, as the “10% definition” still reflects the affordability of household heating, which aligns with the fundamental concern underpinning the fuel poverty concept.

2. Section 5.1.1: When calculating the annual fuel cost C_{AF} in Eq. (5.1), this thesis assumes that the “other energy consumption” E_{other} is entirely electricity, with the unit cost set to the standard electricity tariff. Based on household reports for England and Scotland [43,44], “other energy consumption” is primarily attributed to electricity. Hence, this assumption is not expected to significantly affect the results.
3. Section 5.1.1: England’s fuel poverty data categorises heating fuels into three broad types: 1) mains gas, 2) electricity, and 3) other fuel types. For electricity and other fuels, this thesis estimates unit heating costs using weighted average prices of specific heating options, with weights derived from national statistics in other reports. While this simplification may introduce some inaccuracies since there can be regional variance, it is unlikely to alter the relative economic benefits of HPs. However, this assumption may lead to less accurate geographical distribution of the HP benefits.
4. Section 5.1.1: The efficiency values used for heating systems are based on new installations. Existing systems may operate at lower efficiency, but relevant data is limited. This assumption could lead to an underestimation of the economic benefits of HPs.
5. Section 5.1.1: Fixed costs, such as standing charges for metered fuels and maintenance or delivery costs for non-metered fuels, are not considered. However, as discussed, this can be roughly offset by the HP maintenance cost, resulting in a minor impact on the results.
6. Section 5.1.3: The primary substation (PS) dataset is sourced from DNO websites, which have incomplete information. Entries with missing data are dropped, and power factors (PF) for incomplete records are set to the average PF (0.9855) of available entries. Dropping 3.15% of PSs may slightly overestimate the network upgrade cost, as fewer PSs share the increased load. As for PF, 95% of PSs have PFs above 0.95, so the impact of using the average PF is expected to be minor.
7. Section 5.1.3: The number of households per PS is estimated in Eq. (5.7) based on proportionality to the PS’s existing peak demand. This could be inaccurate as peak demand may also reflect industrial loads.
8. Section 5.1.3: The proportion of off-gas and electric heating options for each PS is estimated by combining statistics from the two nearest MSOAs (for England) or IZs (for Scotland), based on our observation that one PS serves approximately two MSOAs/IZs. This may lead to inaccuracies, as some PSs may serve a larger or smaller number of areas (also MSOAs and IZs have different sizes).

9. Section 5.1.3: The electric heating demand profiles for HP, electric storage heating (ESH), and electric direct heating (EDH) are clipped and rescaled according to the heating pattern assumptions used to derive the official household datasets, which are core to our fuel poverty analysis (heaters are operational only from October to May, for 9 hours on weekdays and 16 hours on weekends). This may underestimate the peak demand during these assumed “non-operational periods” (because in reality some heaters can be operational), leading to an underestimation of network upgrade costs.
10. Section 5.1.3: Existing PS demand profiles are estimated by scaling down the UK national electricity demand profile to match PS peak demand. As local demand patterns are more variable than national averages [31], this assumption may underestimate the true network upgrade costs.
11. Section 5.1.3: When scaling PS upgrade costs to estimate the total distribution network cost (Fig. 5.3), this thesis assumes a uniform scaling factor across all DNO regions. This may reduce the regional variability in the estimated network cost geographical plot.

5.2 Results

This section first presents the fuel poverty estimates before the HP replacement, followed by an analysis of the HP impact on fuel poverty and GHG emissions. Next, the network upgrade cost associated with HP deployment are estimated. Finally, a cost-benefit analysis is conducted, integrating the findings from the preceding analyses.

5.2.1 Fuel poverty estimation

Fuel poverty is a pressing issue that affects a substantial portion of the UK population. Households experiencing fuel poverty struggle to pay the energy needed to maintain a healthy and comfortable indoor temperature.

Following the method in Section 5.1.1, this section first shows the estimated fuel poverty rates for Scotland and England as presented in Table 5.1 (‘Before HP’). The findings indicate that the number of fuel-poor households has nearly doubled due to the energy crisis. Fig. 5.1 offers a regional perspective on the estimated fuel poverty rates, demonstrating that some areas are especially susceptible to fuel poverty. Note that Scotland’s regions show a higher variation in fuel poverty rates and some regions see higher fuel poverty rates than England, which may be attributed to the lower temperature in Scotland, the finer geographical resolution for Scotland’s results (32 regions compared to 9 regions in England), and the higher proportion off-gas households as shown in Fig. 5.1c. Also, while there have been government estimations of fuel poverty at a more granular regional level [54], this chapter conducts the own estimations to ensure consistency with the subsequent fuel poverty analysis using the own scenarios, i.e., fuel poverty rates when replacing off-gas heating with HPs in both pre-crisis and crisis

Table 5.1: Fuel Poverty Rate (%) Estimation

	Scotland		England	
	Before HP	After HP	Before HP	After HP
Pre-crisis	35.55	33.06	17.38	16.67
Crisis	67.26	65.16	47.68	46.85

periods. Such analysis requires household-level datasets. The publicly available versions of these do not disclose household geographical information lower than the government region level for England and the LoA level for Scotland as Fig. 5.1. Throughout this chapter, 2019 prices are considered pre-crisis levels, and 2022 prices are utilised for estimation during the energy crisis.

To delve deeper into the fuel poverty issue, this chapter explores the proportion of off-gas households in each region, as illustrated in Fig. 5.1c. Intriguingly, the fuel poverty rate heatmap exhibits a pattern similar to the off-gas proportions. The Pearson correlation coefficient between the off-gas proportion and the pre-crisis (resp. over crisis) fuel poverty estimation is 0.81 (resp. 0.72). The primary reason could be the higher cost of off-gas heating options such as ERH, heating oil, and LPG. Note that even though HPs are a form of electric heating, as of the 2022 winter, only 1% of UK households were using HPs as their primary heating method [352]; hence this chapter assumes a baseline HP adoption rate of zero. This finding is also reported in the official fuel poverty reports in England [54] and Scotland [43] respectively. Consequently, eco-friendly and economically viable heating technologies are of utmost importance for off-gas households.

5.2.2 Fuel poverty reduction by heat pumps

HPs are recognised as a cost-effective heating option due to their high heating efficiency. The HP efficiency varies across a year in response to the change in ambient temperature. As the fuel poverty evaluation focuses on the annual energy bill, the seasonal performance factor (SPF) is a better metric to measure the overall efficiency of HPs over a year. Different sources have reported different estimations of the SPF for HPs. The UK Renewable Heat Incentive (RHI) scheme has an average design SPF of 300% for air-source HPs in 2018 [353], which increases to 360% for new installations in 2022 [354]. A recent HP field trial observed an SPF at 280% for air-source HPs [355], showing a significant increase compared to past field trials and a good performance even for cold days. Based on these data, it is expected that the coming HP rollout can reach an SPF at 300% for both the England's and Scotland's regions in the subsequent case studies. Furthermore, the later focus of this chapter

Table 5.2: Fuel Poverty Rate by Region

Nation	Region	Samples	FP Pre-crisis (%)	FP During Crisis (%)
Scotland	Clackmannanshire	219	30.26	61.85
Scotland	Dumfries and Galloway	266	53.74	81.07
Scotland	East Ayrshire	250	41.37	78.81
Scotland	East Lothian	232	33.01	63.87
Scotland	East Renfrewshire	225	27.13	64.84
Scotland	Na h-Eileanan Siar	271	72.70	94.71
Scotland	Falkirk	247	33.97	61.88
Scotland	Highland	257	58.47	83.02
Scotland	Inverclyde	214	38.03	65.76
Scotland	Midlothian	273	30.67	61.11
Scotland	Moray	199	56.54	84.81
Scotland	North Ayrshire	221	46.98	74.44
Scotland	Orkney Islands	263	65.66	89.99
Scotland	Scottish Borders	208	47.07	78.98
Scotland	Shetland Islands	273	70.53	92.24
Scotland	South Ayrshire	228	46.01	73.11
Scotland	South Lanarkshire	325	33.01	67.30
Scotland	Stirling	237	34.86	65.69
Scotland	Aberdeen City	240	37.78	60.24
Scotland	Aberdeenshire	242	48.15	78.43
Scotland	Argyll and Bute	229	52.00	83.53
Scotland	City of Edinburgh	598	27.39	60.16
Scotland	Renfrewshire	188	27.77	58.89
Scotland	West Dunbartonshire	223	38.67	72.11
Scotland	West Lothian	263	23.89	53.59
Scotland	Angus	256	40.37	78.14
Scotland	Dundee City	255	43.64	74.08
Scotland	East Dunbartonshire	245	28.40	62.48
Scotland	Fife	439	35.68	71.41
Scotland	Perth and Kinross	262	43.95	72.03
Scotland	Glasgow City	621	31.20	64.66
Scotland	North Lanarkshire	333	27.37	62.51
England	North East	729	19.40	54.48
England	North West	1659	17.54	50.95
England	Yorkshire and The Humber	1378	18.62	52.23
England	East Midlands	1074	14.39	43.66
England	West Midlands	1088	20.76	52.36
England	East of England	1446	17.87	47.59
England	London	1441	19.11	43.90
England	South East	1995	13.67	41.83
England	South West	1164	16.90	49.00

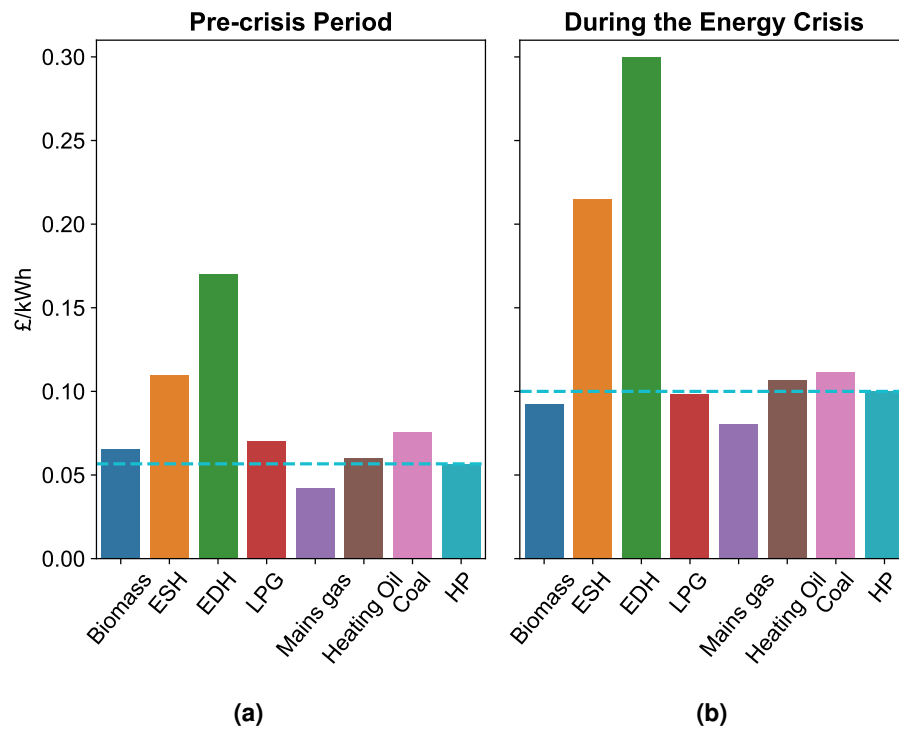


Figure 5.4: Unit heating costs of different heating options.

(a) Unit costs before the energy crisis. (b) Unit costs during the energy crisis. The dotted line is the HP cost and is included to make comparison more convenient. Data for electricity, gas mains, heating oil, and coal prices are obtained from BEIS [337, 338]. The prices of LPG and biofuels are based on Nottenergy [339]. The detailed derivation of the unit heating cost is given in Section 5.1.1.

will be on air-source HPs only, due to their broader applicability than ground-source HPs as discussed in [52] (ground-source HPs have stricter installation requirements and are more expensive, despite having higher efficiency). The UK future energy scenario also holds the same expectation [180].

Fig. 5.4 compares the unit heating cost of HPs with other heating options during the pre-crisis and crisis periods, following the method detailed in Section 5.1.1. Here the unit heating cost is the cost per unit of heating energy, which is calculated by the method in Section 5.1.1. As illustrated, HPs have a lower running cost than all of the off-gas heating during the pre-crisis period, particularly compared to EDH and ESH. During the energy crisis, HP becomes less competitive than 'Biomass' and 'LPG', but the HP running cost is still lower than 'ESH', 'EDH', 'Heating Oil', and 'Coal', which account for around 90% off-gas heating in England and Scotland, calculated from the census data [340, 346].

Upfront cost

The low running cost of HPs makes it possible to see a reduced fuel poverty rate if off-gas heating is replaced with HPs. However, HPs also have a much higher upfront cost than other heating options. For a typical semi-detached UK household, the upfront cost for an air-source HP without the available subsidy is £8,350, which covers the cost of the HP unit, installation, a hot water cylinder, and the necessary installation of low-temperature emitters [53] (note that For HPs, there is also an option to replace the entire heating distribution system rather than just upgrading the emitter, which may lead to £6,000 more cost suggested by [356]). The complete replacement can reduce the time to heat up the rooms [356]; however, this is not considered a fundamental heating need and is thus not included in the fuel poverty evaluation of this chapter). In comparison, the upfront cost for a combination gas boiler, including the boiler and installation, is only around £1,570-£2,250 [53, 356, 357]. An ESH comes with an upfront cost of £4,130, while an EDH costs around £2,500, covering the heater, installation, and hot water cylinder [53]. A combination oil boiler has an upfront cost ranging from £1,791 to £3,560, which includes the cost of the boiler and installation [356, 357]. All of these heating methods have lower upfront costs than HPs.

The high upfront costs of HPs can dissuade residents and landlords from adopting HPs despite their low running costs. Therefore, considering the upfront cost is crucial when analysing the economic benefits of HPs. To tackle this issue, amortisation with a certain interest rate can be configured to distribute the cost across HP life (see Section 5.1.1). The life expectancy of HP ranges from 15-25 years, according to [358, 359]. This chapter sets the life expectancy to 20 years as an intermediate value. Note that this thesis uses the annual payment towards amortisation for its better intuitiveness; however, this is equivalent to the levelised energy cost method, as explained in Appendix B.3. Fig. 5.5 summarises the fuel poverty reduction that considers the HP upfront cost. It can be seen that even under a low interest rate of 3%, there are still 30% of regions with increased fuel poverty rates during the pre-crisis period (first row). More regions experience reduced fuel poverty rates during the energy crisis (second row) because the reduction in running costs by HPs becomes more significant than their upfront costs during the crisis period. However, under a normal 20-year secured homeowner loan interest rate by a UK loan company (around 7%) [360], there are still $1-65.9\%=34.1\%$ regions with increased fuel poverty rates after HP replacement.

The England and Scotland governments have policy support for HP uptake. In England (also Wales), the Boiler Upgrade Scheme rewards up to £5,000 (increases to £7,500 from October 2023 [361]) for each household installing an air-source HP. In Scotland, the grant can cover 75% of the costs up to a maximum of £7,500, with the remainder taken as an optional interest-free loan [362]. The second column in Fig. 5.5 considers the England government's £5,000 support to the HP upfront cost. A significant uplift in fuel poverty reduction rate is observed, with at least 90.2% regions seeing reduced fuel poverty rates during the pre-crisis period

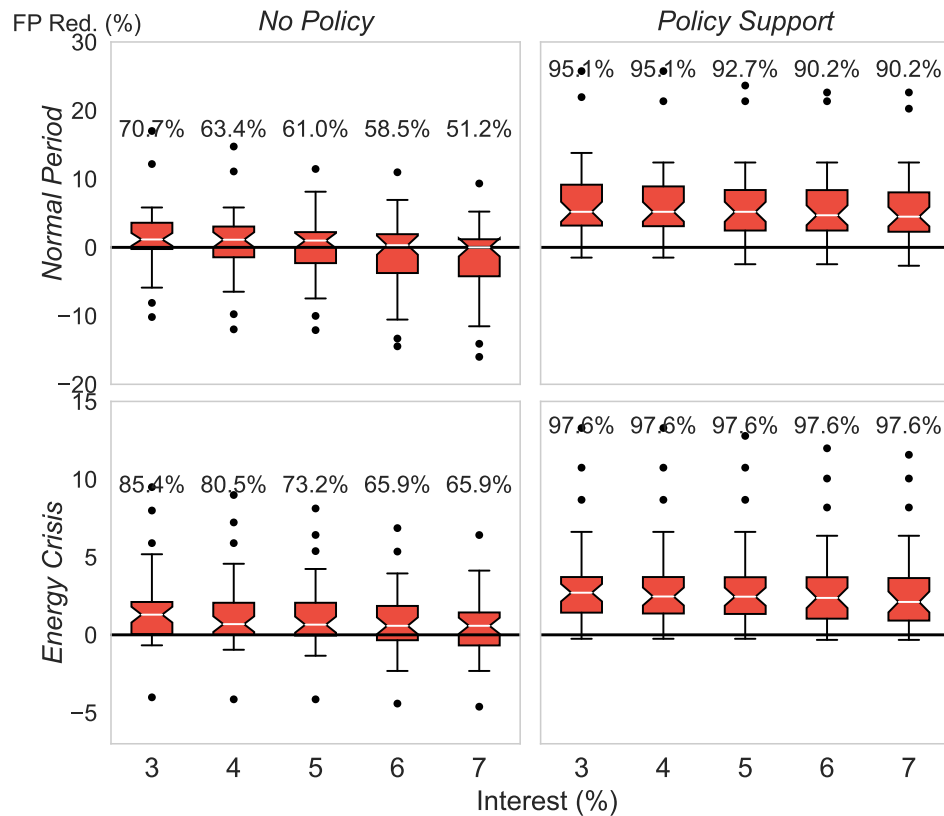


Figure 5.5: Proportion of out-of-fuel-poverty households considering the HP upfront costs.

Each sample corresponds to the result for a England's/Scotland's region. More specifically, each sample represents the regional proportion of households lifted out of fuel poverty (FP Red.) after HP replacing off-gas heating. The numbers on top of each box represent the proportion of samples (regions) that see a positive reduction in the fuel poverty rate. Each box in Fig. 5.5 corresponds to a different amortisation interest rate. 'Policy Support' means that England's £5,000 grant is deducted from the HP upfront cost for the second column. The numbers on top of the boxes indicate the proportion of regions seeing positive fuel poverty reductions, i.e., the proportion of samples above the black horizontal lines.

and 97.6% during the energy crisis, for all interest rates. Also, without policy support (the left column), there are outliers below the boxes in the plot, representing regions with high increases in the fuel poverty rate after HP replacement. With policy support, the outliers below the boxes are eliminated. These results demonstrate the necessity of government support in HP uptake.

The later content of this chapter will analyse the benefits of HPs replacing all the off-gas heating under the existing £5,000 England government support and assume that the household amortises the remaining HP upfront cost for 20 years. The interest rate is set to 7%, which is close to the value provided by a UK loan company [360].

Geographical illustration

This section first evaluates the impact of the HP rollout on fuel poverty rates during the pre-crisis period. Fig. 5.6a illustrates the results at a regional level when replacing off-gas heating with HPs. The reduction rate can be higher than 20% for specific regions.

Then, this section estimates the fuel poverty reduction amid the energy crisis after the HP roll-out. Table 5.1 and Fig. 5.6b show that the fuel poverty reduction brought by HPs becomes less significant during the crisis period. The reduction rate is only at most 11% for specific regions. The reduced effect of HPs during the crisis is because of that the fuel poverty is defined as spending above a threshold. Therefore, when a customer spends far above the threshold, even substantial reductions in their energy bills may still leave them in fuel poverty. To confirm this argument, Figs. 5.6c and 5.6d visualise the percentage fuel cost reduction for off-gas households after HP replacement. Compared to the pre-crisis period, the fuel cost reduction is indeed higher in the crisis period. An obvious lift is observed for the England regions, because in a high-energy-price (crisis) period, the fuel cost reduction becomes more dominant than the upfront cost component. This also means that the HP rollout indeed has a greater impact in terms of bill reduction for customers in fuel poverty, even though the impact on the fuel poverty rate is less apparent.

GHG emission reduction

Another key co-benefit of HP replacement is reduced GHG emissions in the off-gas domestic heating sector due to the lower energy consumption. This section evaluates the proportion of GHG emissions that could be reduced by replacing off-gas heating with HPs. Under the GHG emission factors in the UK in 2020 [345], GHG emissions are approximately 12.44 MT CO₂e in the off-gas domestic heating sector for England and Scotland in total. When off-gas heating is replaced with HPs, GHG emissions drop to 3.86 MT CO₂e, a 69% reduction.

Fig. 5.6e provides a region-level plot of GHG emission reductions, revealing that some regions could experience a reduction above 70%. Furthermore, the majority of regions (92.7%) are expected to see GHG emission reductions of over 60%. One should note that this chapter only considers GHG emissions in the off-gas domestic heating sector, but the on-gas heating sector (i.e. using the natural gas network) still generates emissions. Full decarbonisation of heating will require a large-scale transition, such as replacing natural gas heating with HPs. Finally, although the carbon benefits of HPs are well established in existing literature, the analysis presented in this chapter extends this understanding by quantifying the extent of regional emission reductions specifically through the replacement of off-gas heating with HPs; this assessment is conducted within the context of the current regional off-gas heating

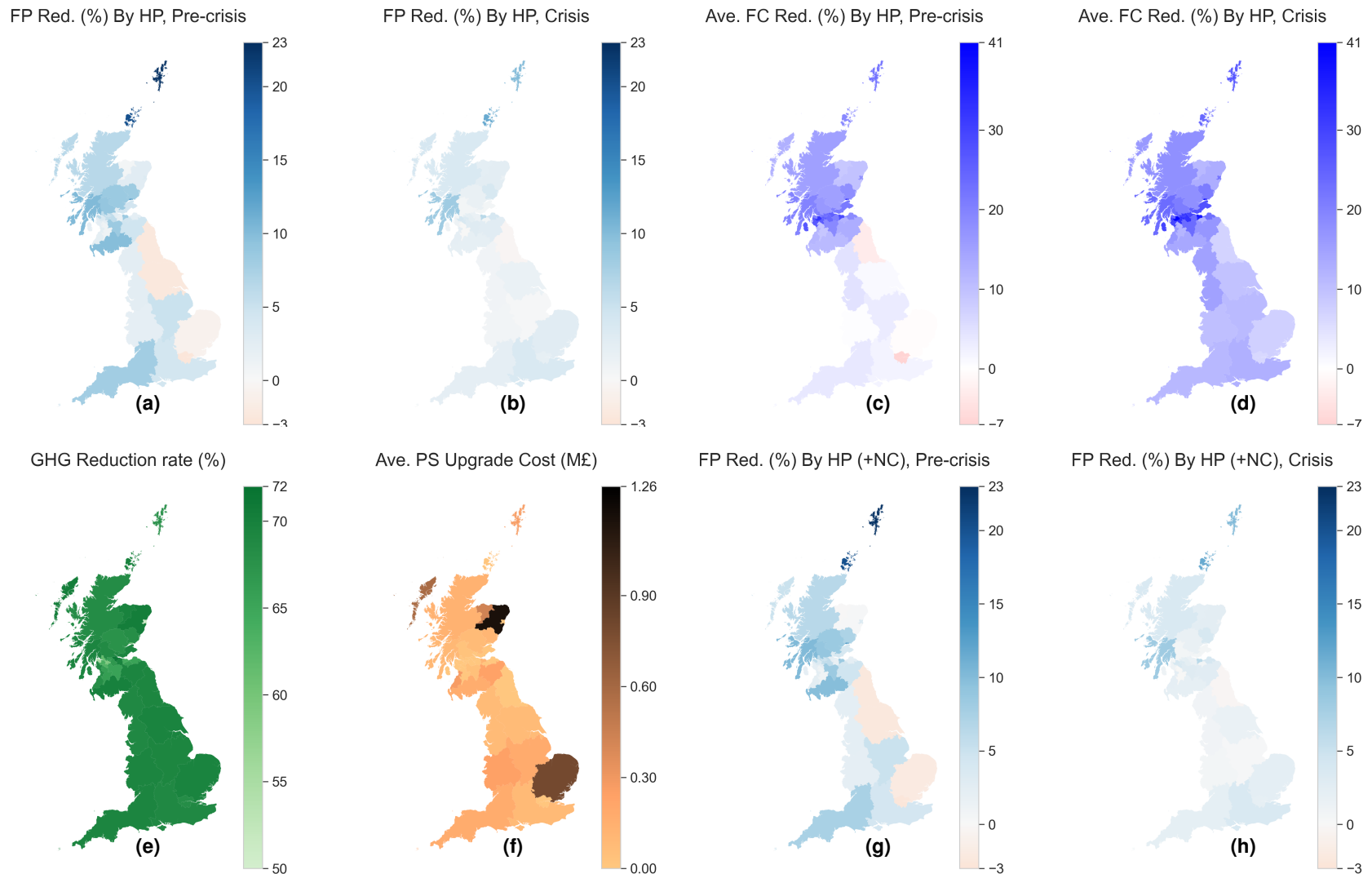


Figure 5.6: Regional fuel poverty reduction, fuel cost reduction, GHG reduction, and network upgrade cost when HPs replace off-gas heating.

(a) Estimated proportion of households out of fuel poverty during the pre-crisis period. (b) Estimated proportion of households out of fuel poverty during the crisis. (c) Estimated average off-gas household fuel cost reduction (FC Red) during the pre-crisis period. (d) Estimated average off-gas household fuel cost reduction (FC Red) during the crisis. (e) GHG reduction rates by region. (f) Average distribution network upgrade cost by region. (g) Estimated proportion of households out of fuel poverty considering the increased energy bills due to the network upgrade cost (highlighted as +NC) based on fuel prices pre-crisis. (h) Estimated proportion of households out of fuel poverty considering the increased energy bills due to the network upgrade cost (highlighted as +NC) based on fuel prices during the energy crisis.

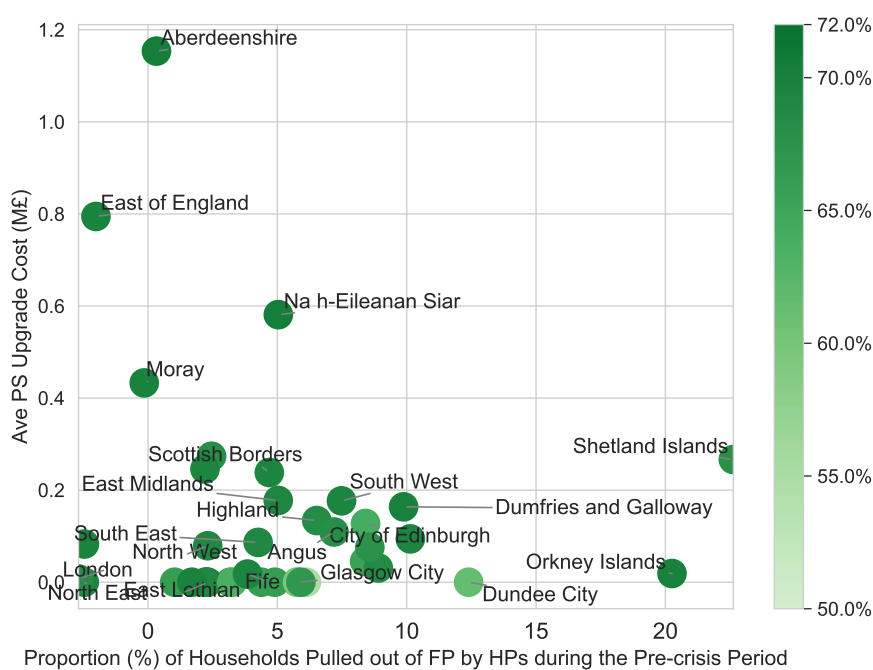
landscape and the corresponding GHG emission factors in the UK. By consolidating the environmental benefits and the previously demonstrated fuel poverty benefits, more comprehensive guidance can be provided to policymakers and other stakeholders who are evaluating the promotion of HPs.

5.2.3 Network upgrade cost

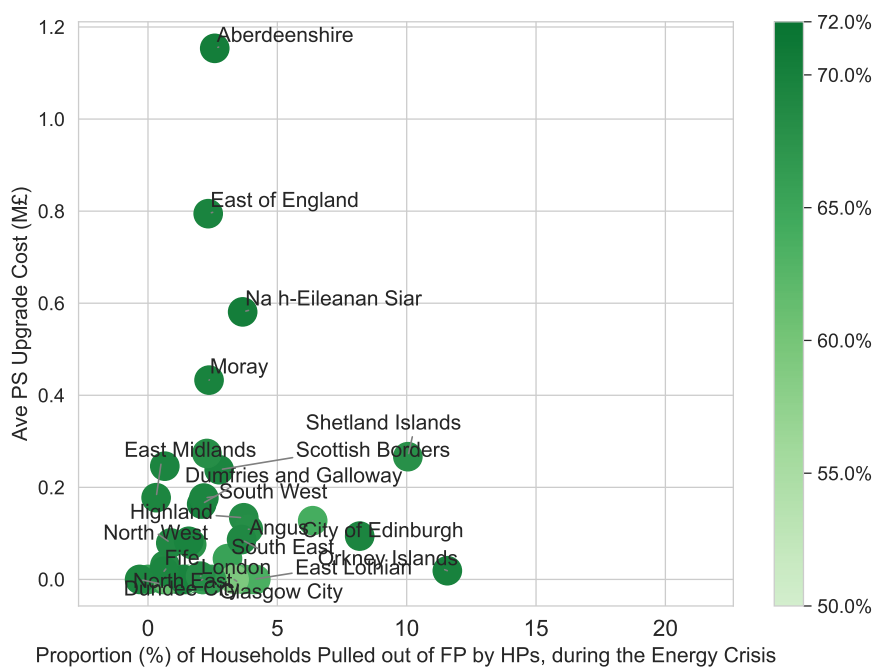
The widespread adoption of electric HPs necessitates a reliable electricity distribution system. If the electricity demand becomes too high, the distribution system may require upgrades, which would eventually result in higher customer bills or government spending. The case is particularly severe for regions with weak local distribution infrastructure and high proportions of off-gas homes when replacing all the off-gas heating with HPs. Only 1% households used HPs as the main heating method in the winter of 2022 [352] and this chapter hereby assumes a zero HP adoption rate in the current off-gas heaters.

The network upgrade cost is estimated based on the proposed method in Section 5.1.3. This chapter estimates that 457 out of 3,891 PSs in England and Scotland will require upgrades to support the additional HP demand. The total distribution network upgrade cost is estimated to be £715.4 M. For a more detailed breakdown of regional costs, Fig. 5.6f shows the average upgrade cost (the total cost divided by the number of PSs in each region) for all regions of England and Scotland. As can be seen, Aberdeenshire in the top-right of the map has the highest average network upgrade cost due to its relatively high proportion of off-gas homes (41.2%). East of England in the bottom-right part is the second highest, although its proportion of off-gas homes in Fig. 5.1c is relatively low (17.8%)—this could be attributed to the low existing demand headroom in their distribution network, so even a small HP rollout can trigger a high cost. In particular, the two archipelagoes in the north (the Orkney and Shetland islands) with almost 100% off-gas homes still see relatively low upgrade costs. This counterintuitive result comes from the fact that more than half of households have used electric heating and replacing these electric heaters with efficient HPs lead to less stress on the network. The high network capacity-to-population density ratio could be the other reason.

The network upgrade cost will finally be repaid by customers through their increased energy bills [351], which in turn affects the fuel poverty results. To consider this impact, this chapter allocates the regional network upgrade cost to households over a 45-year period, matching the equipment lifespan [363], and ensuring that the net present values of additional energy bills over this period equal the total upgrade cost (see Appendix B.4 for details). This chapter illustrates the updated fuel poverty reduction results in Figs. 5.6g and 5.6h. It can be seen that the HP rollout still brings substantial fuel poverty reductions to most regions after considering the additional repayments towards the network upgrades. Comparing Figs. 5.6a and 5.6g, reduced fuel poverty benefits are observed over the pre-crisis period in areas with the highest network costs, specifically Aberdeenshire (reducing from 2.75% to 0.33%) and the East of



(a)



(b)

Figure 5.7: Fuel poverty (FP) reduction and GHG emission reduction versus network upgrade cost.

(a) The plot during the pre-crisis period. (b) The plot during the energy crisis. The colours of each plot represent the GHG emission reduction rate.

England (reducing from -0.84% to -2.01%), where the latter result suggests an increase in households experiencing fuel poverty after the replacement of off-gas heaters by HPs when network costs are considered. During the crisis period (comparing Figs. 5.6b and 5.6h), there is a minimal difference, as additional energy costs due to network upgrades become less impactful given the already high energy bills during the crisis. However, the reduced benefits in the pre-crisis period underscore the non-trivial impact of network upgrade cost on the long-term benefits of HPs for specific regions.

The next price control period RIIO-ED2 (from 2023 to 2028) also estimated the total distribution network upgrade cost to be £3.2 B to accommodate EVs, HPs, and local low-carbon generations [351]. This is higher than the estimated £715.4 M of this chapter because 1) additional HP demand in the on-gas properties is not modelled here considering the lower economic competitiveness of HP running cost compared to mains gas, 2) this chapter replaces all the existing electric heaters in the off-gas sector with HPs, reducing the electric demand due to HP's high efficiency, and 3) this chapter separates the impact of HPs given the existing distribution network headroom, separating it from network upgrades due to parallel developments like the adoption of EVs. This separation aligns with the specific focus of this chapter on assessing the benefits and costs associated with HPs. By keeping these factors separate, this chapter can also attribute the additional energy cost and consequent effect on fuel poverty directly to HPs, without mixing in other contributing elements. In reality, the total network upgrades will also be triggered by other system changes, such as EV adoption and local renewable generation, which will overlap with the upgrade costs for HPs replacing off-gas heating.

5.2.4 Cost-benefit analysis

This section performs a combined cost-benefit analysis of HPs to provide more regional insights.

Maximising benefits while minimising costs

Fig. 5.7a depicts the potential pre-crisis benefits of HP deployment, alongside the associated network upgrade cost. The fuel poverty benefits have considered the repayments towards the network upgrade cost. In the figure, regions closer to the right-bottom corner and marked by darker colours have the highest benefits (reducing fuel poverty and GHG emissions) and lowest network costs associated with HP deployment. Fig. 5.7a reveals that most regions have relatively low network upgrade cost. In particular, certain regions such as Shetland and the Orkney Islands exhibit low cost and relatively high benefits, making them good candidates

for government-led HP pilot projects and deployment initiatives. In contrast, few regions, such as Aberdeenshire and East of England, see fewer benefits and high network cost. These results can help policymakers design and prioritise regional HP support mechanisms and deployments.

Fig. 5.7b shows the cost and benefits during the energy crisis period. The majority of regions see a shift to the left side as a result of the high energy expenditure during the crisis. However, their relative positions, such as the most “prior” regions (Shetland and the Orkney Islands) and less “prior” regions (Aberdeenshire and East of England), persist overall. In summary, a cost-benefit analysis can inform the government decision-making process regarding HP deployment, ensuring efficient and effective resource allocation to accomplish both environmental and economic objectives.

Multi-dimension correlation analysis

Fig. 5.8 presents an extensive correlation graph, examining the relationships among the key dimensions studied in this chapter. The lower triangle plot is based on the Pearson correlation for linear relationships, while the upper triangle is based on Spearman’s coefficient, capturing nonlinear relationships. The plot indicates that regions with high fuel poverty rates are highly likely to experience even higher rates during the energy crisis, motivating policy interventions in these vulnerable areas. The off-gas proportion exhibits a strong positive correlation with both the fuel poverty rate and the fuel poverty reduction by HPs during both periods. This result demonstrates the vulnerability of highly off-gas regions but also a stronger positive effect by HP roll-out, which may be worth a prioritised deployment for more affordable and sustainable heating alternatives, such as HPs. The high linear correlation (0.7) between fuel poverty reduction by HPs in pre-crisis and crisis periods also suggests that HPs could be a robust solution for mitigating fuel poverty.

This chapter does not observe a linear correlation (close to 0) between network upgrade cost and fuel poverty reduction in both periods. The Spearman’s correlation (0.30) suggests a positive but weak monotonic relationship in the crisis period. This result indicates that higher economic benefits may not necessarily come with higher network upgrade cost. Indeed, regions such as Shetland and Orkney Islands enjoy a high reduction in fuel poverty and low network upgrade cost, as discussed in Sections 5.2.4 and 5.2.3. Although there is a positive Spearman correlation (0.55) between the network upgrade cost and GHG reduction rate, the variation in GHG reduction across regions, as depicted in Fig. 5.6e, is not substantial. These findings highlight that the higher cost is not necessarily attached to the higher benefits of HP replacement at the regional level.

Fig. 5.8 also analyses the correlation for the proportion of existing electric heating before the HP rollout ('elec proportion'). Specifically, the 'elec proportion' still exhibits positive Spearman's correlation (0.52) with the network upgrade cost ('Ave Upg. Cost'). Although replacing existing electric heating with HPs can reduce the network demand so as the network upgrade cost, Fig. 5.8 shows that the 'elec proportion' is positively correlated with the 'offgas proportion' (Pearson correlation of 0.96 and Spearman's correlation of 0.93)—therefore, the reduction of network demand by the increasing 'elec proportion' is not sufficient to offset the increase in network demand due to the associated higher proportion of off-gas households. However, if we observe the proportion of existing electric heating among off-gas households ('elec in offgas prop'), the correlation with the network upgrade cost is negative (Pearson correlation of -0.47 and Spearman's correlation of -0.52). This aligns with the intuition that replacing these electric heaters with HPs can lead to a lower demand on the distribution network.

Finally, it is worth noting that the correlation analysis in this chapter is based on the regional data, which may not be representative of individual households. The results should be interpreted with caution, as they may not apply to specific cases or smaller geographical areas.

It is essential to acknowledge the distinction between correlation and causation. The analysis in this chapter primarily identifies the correlation among different dimensions but does not establish causality. More specific analysis and experiments are needed to interpret the causal relationship between various factors [364].

5.3 Discussion

HPs have a low running cost but high upfront cost. Taking the upfront cost into consideration leads to even higher fuel poverty rates for many regions when off-gas heating is replaced with HPs. Government support is necessary, and this chapter shows that the economic benefits of HP replacement can be restored with the England government grant. Meanwhile, with the massive production of HPs in the future, HP may enjoy a lower upfront cost, making it more economically viable. Another approach to increase the economic benefits of HPs can be the integration of local renewable energy sources, such as local solar or wind generation. An online survey carried out in the UK [365] shows that only 33.2% households would not consider installing solar panels at their home, demonstrating the feasibility of local RES deployment. Private-owned generation may alleviate the electricity cost and therefore the cost of HP heating. Local RES can also be installed for an entire community or shared among households to reduce the upfront cost, which can be incentivised by an appropriate local energy market (LEM) design that fairly distributes cost and profits [4, 366].

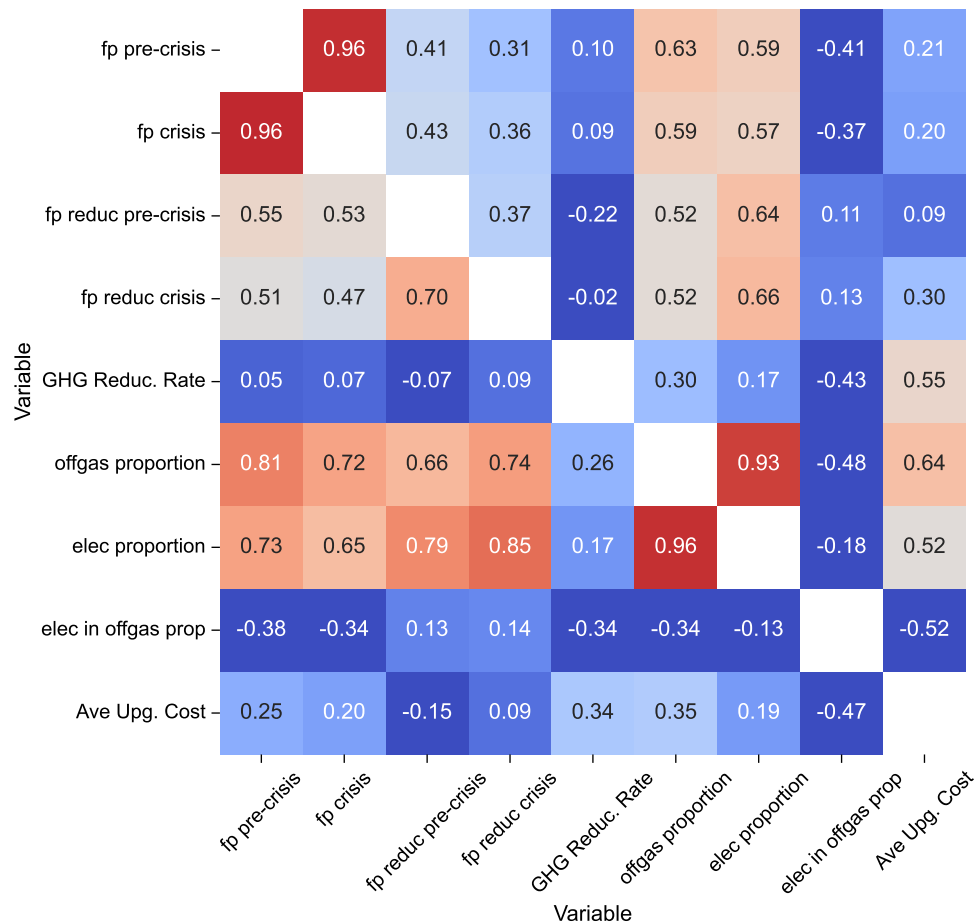


Figure 5.8: Correlation plot for the central dimensions for the analysis of HPs on fuel poverty, network upgrade costs, and GHG emissions.

“fp pre-crisis” is the fuel poverty estimate over the pre-crisis period. “fp crisis” is the fuel poverty estimate over the crisis period. “fp reduc pre-crisis” is the fuel poverty reduction over the pre-crisis period after the HP replacement. “fp reduc crisis” is the fuel poverty reduction over the crisis period after the HP replacement. “GHG Reduc Rate” is the GHG emission reduction rate after the HP replacement. “offgas proportion” is the regional proportion of off-gas households. “elec proportion” is the regional proportion of households with electric heating (direct and storage) before HP rollout, “elec in offgas prop” is the proportion of existing electric heating (prior to the HP rollout) among off-gas heating options. “Ave Upg Cost” is the average distribution network upgrade cost for accommodating the HP replacement. The lower triangle plot is based on Pearson correlation for a linear relationship while the upper triangle is based on Spearman’s coefficient that captures a nonlinear relationship.

A recent study [52] concluded a negative impact on fuel costs when replacing the mains gas heating with HPs, even without considering the upfront cost of HPs. Fig. 5.4 has a similar observation that HPs have a higher running cost than the mains gas. The case studies in this chapter illustrate that the economic benefits of HPs hold for off-gas households.

In addition to the £5,000 (increases to £7,500 from October 2023 [361]) government subsidy for HP installations, the UK has introduced several policies to address fuel poverty, such as the Energy Bills Support Scheme, Energy Price Guarantee, and subsidies for energy efficiency measures. A valuable avenue for future research is to determine which of these policies has the most substantial impact on reducing fuel poverty. Addressing this question can guide the development of more effective policies for alleviating fuel poverty using public funds. A related approach was demonstrated in a study by Ref. [367], where the fuel poverty metric was used to plan targeted energy interventions.

HPs can be flexible in terms of the ability to be turned on/off for a short period while maintaining residents' comfort. This flexibility makes HP capable of engaging in the local flexibility market (LFM). LFM is an emerging technology that UK DNOs are adopting [368], where energy flexibility can be traded to provide services to the power system and bring profits to end-users [369]. The previously mentioned LEM and LFM here with different time horizons can also be stacked for HP engagement. The Local Energy Oxfordshire (LEO) project is considering and evaluating the market value stack in the Oxfordshire area of the UK [370]. LEO is also testing the flexibility provision of smartened HPs in Oxfordshire residents [371]. However, it is also worth noting that HPs may have smaller flexibility than other heating options: the high upfront cost of HPs makes the installed thermal power usually smaller than other heating options [356], leading to a slower indoor temperature adjustment.

In addition to economic barriers, air source HP installation can also cause significant disruption due to the need to install a bulky fan unit [372], which may be less acceptable to some residents. Finally, for large-scale HP roll-out, there could be insufficient HP production or installers. If these supply chain issues persist, the fuel poverty reduction brought about by HPs during the energy crisis period may not appear as evident, as the natural gas price is projected in [373] to return to normal by 2030. Nevertheless, geopolitical tensions and uncertainties indicate that gas prices will continue to be volatile, arguably more so than the increasingly renewable-intensive electricity [374]. A range of policies is also now in place to address these issues and accelerate HP deployment. These policies include the Boiler Upgrade Scheme to attract public investment in HPs, the "Heat Pump Investment Accelerator Competition" to support component manufacturing infrastructure across the UK, and the "Heat Training Grant" to increase the number of qualified HP installers, etc [375, 376]. It is expected that these policy measures can, to some extent, expedite the HP rollout process. Furthermore, while the benefits are contingent on supply chain and installation processes, the initial step is to acknowledge the potential benefits that HPs can offer to address the fuel poverty issue,

which is the primary focus of this chapter. Lastly, this chapter demonstrated that HP benefits endure in both normal and crisis periods. This means that even if the crisis ends before HP adoption reaches a substantial level of impact, the adoption of HPs can still yield long-term benefits, particularly considering the possibility of future energy crises.

Building efficiency is also important for HP deployment. Although building fabric is less likely to affect the HP efficiency [377], poor insulation will require larger-size HPs, resulting in higher upfront costs. Rural regions are likely to have this building efficiency issue, hindering the financial benefits of HPs. However, similar to the capital costs of solar and wind energy which have fallen 40% and 80% respectively during the past 10 years [52], it is anticipated to see a lower upfront cost of HPs with the development of HP markets, thereby alleviating the additional HP upfront cost in less-efficient buildings.

The wide-range HP roll-out can threaten the electric distribution network as analysed in Section 5.2.3. As an alternative option for heating electrification, ESH has the advantage of being able to shift demand to off-peak periods, thus alleviating the burden on the power grid. HPs can also be equipped with additional thermal storage, alleviating the burden on the power grid like ESH and also saving energy bills by charging on a low-tariff period [372]. Storage can also offer HPs with higher demand flexibility, which facilitates more stable power system operation and brings more economic benefits to HP owners by participating in demand response and various LEM and LFM [378, 379]. The integrated HP and storage system is recognised as a key to consumer demand satisfaction and a stable electricity grid by the International Energy Agency [380].

Note that HP is not always a burden on the power grid. For regions where ERH is being massively adopted, replacing the ERH with HPs can even lead to lower electric demand because of the high efficiency of HPs. Section 5.2.3 has already illustrated this effect, where the two off-gas-grid archipelagoes (the Orkney and Shetland Islands) with high ERH adoption enjoy a low upgrade cost after HP replacement.

The distribution network upgrade cost in this chapter is scaled from the PS level due to the limited data availability. The more precise analysis would require detailed network information that is protected by the corresponding DNOs.

The rebound effect is an extensively discussed fact that could lead to actual energy saving lower than expected [381]. One example is that households may consume more energy than before after the energy efficiency is improved. This effect may lead to less energy consumption reduction and less GHG emission reduction by HPs than simulated. The current fuel poverty definition in the UK seems to avoid this issue as the fuel poverty indices are calculated based

on the modelled energy consumption, rather than those actually consumed. However, the peak electricity demand and consequently network upgrade costs could increase due to the rebound effect. This may in turn raise modelled fuel costs through higher electricity prices needed to recover DNOs' investments, thereby reducing the fuel poverty benefits.

The UK has significant socio-economic inequalities, especially between the north and south [52]. For example, it is shown that productivity, measured as income divided by working hours, is on average lower in the northern regions than southern regions. The pandemic has exacerbated this issue. This chapter has illustrated that the benefits and costs brought by HPs are not evenly distributed among the Scotland and England regions. Consequently, it is crucial to exercise prudence and prioritise equitable outcomes during the HP deployment process, aligning with the core objectives of the UK Levelling Up programme [382], which seeks to address existing inequalities.

5.4 Limitations of study

The findings in this chapter have some limitations. First, this chapter applied the 10% definition of fuel poverty. Scotland's metric additionally considered the remaining income after deducting the fuel costs, and England's definition includes the "low fuel poverty energy efficiency rating". Some energy bill discounts, such as the Warm Home Discount used in the official fuel poverty estimations, were not considered in this chapter. Therefore, the estimated fuel poverty results could deviate from the actual official fuel poverty estimations. Justifications for the 10% definition and a detailed comparison with the current official metrics in the UK are presented in Section 5.1.1. Future work can adjust the metrics and make new definition through the process in Section 5.1.1 with necessary data. Note that the economic benefits of HPs for off-gas homes should persist as the energy cost reduction is independent on the fuel poverty definition.

Second, the estimated distribution network cost serves as an indicator of the potential network upgrades and could also deviate from the actual network upgrade cost. Accurately estimating network upgrade cost is complicated and requires more information regarding the distribution network.

Finally, as discussed in Section 5.1.4, this chapter made several assumptions in the modelling and estimation process, driven by data unavailability and the complexity of the problem. As more granular datasets become available, future work could refine the analysis and deliver more accurate results.

5.5 Concluding remarks

This chapter analysed the impact of HP adoption on fuel poverty and electricity distribution networks when replacing expensive off-gas heating in England and Scotland. Regional fuel poverty rates and the associated GHG emission reductions are evaluated. The analysis demonstrated positive effects under both pre-crisis and crisis conditions, assuming existing government support for HP upfront costs.

The correlation analysis in this chapter revealed that high fuel poverty benefits do not necessarily correspond to high network upgrade cost at the regional level. This chapter identified “priority” regions that exhibit low costs and high benefits, presenting targeted opportunities for policymakers. However, certain regions may experience significant burdens on electricity distribution networks.

To ensure that the fuel poverty reduction benefits of HPs are equally extended across the entire UK, addressing the network challenges in these high-stress regions becomes critical; otherwise, high network upgrade cost will be triggered, which will finally increase the customers’ energy bill as discussed in Section 5.2.3. This underscores the importance of activating HP flexibility, which can help alleviate network stress and support more equitable deployment across all regions. Analysing the benefits and feasibility of activating such flexibility can be a valuable direction for future research.

This chapter addressed the second research question by evaluating the impact of HPs on fuel poverty and distribution network costs when HPs are used to replace off-gas heating. In doing so, this chapter tackled challenges that complement technical barriers in unlocking DER flexibility studied in previous chapters. Understanding the impacts of DERs is equally significant to addressing technical challenges, as insufficient insight can impede strategic planning and delay the realisation of DER flexibility benefits. In fact, in addition HPs, data centres represents another demanding but underexplored DERs, which is the focus of the next chapter.

Evaluating the Grid Flexibility and the Associated Cost of Data Centres

As discussed in Chapter 1, understanding the power-system, economic and social impacts of DER also represents a crucial dimension that supports the design of effective and efficient roadmaps to realise DER flexibility. However, two major types of DERs, namely HPs and data centres, remain under-investigated in existing studies. The previous Chapter 5 addressed this challenge in terms of the under-investigated fuel poverty and network impact of HPs, a major DER for heat electrification. This chapter turns its focus to another important yet similarly underexplored category of DERs, namely data centres, and seeks to evaluate and understand their potential in providing power system flexibility. This evaluation also answers the last research question of this thesis:

Q3: Do AI-focused data centres have a different potential for grid flexibility compared with general-purpose data centres, and what is their cost of flexibility provision?

As discussed in Sections 1.1.3 and 2.3, data centres represent a substantial source of electricity demand, which is exacerbated recently due to the surge in AI. According to the EPRI, data centres currently account for 4% of annual electricity consumption in the US, a figure projected to rise to as much as 9.1% by 2030, driven largely by AI workloads [64]. This growing energy demand from AI presents a significant challenge for power grid infrastructure [65], especially when considered alongside efforts to electrify heating and transportation to achieve net-zero carbon emissions [49].

A promising solution to alleviate the power grid impact created by data centres is to make their power demand flexible, by adjusting computing workloads and consequently power consumption. This flexibility can be utilised to provide power system services that are critical for maintaining normal and stable power system operations [11]. Existing research has explored the potential of traditional CPU-intensive, general-purpose HPC data centres to provide various power system services. However, as illustrated in Section 2.4.2, several key gaps remain:

1. **Underexplored distinct characteristics of AI-focused data centres:** The power consumption and computing patterns of AI-focused HPC data centres can differ significantly from those of traditional general-purpose HPC data centres. These distinct characteristics have not been explored in the context of power system services. Filling this gap can contribute to develop more targeted data centre flexibility schemes tailored to the specific characteristics of different types of data centres.
2. **A lack of cost modelling for data centre flexibility:** There is a lack of models that quantify the cost of data centres providing power grid flexibility. This gap leaves data centre operators disincentivised due to the unclear financial profitability of offering power system services, as these services may create disruptions to valuable computing tasks.
3. **A lack of assessment for various power system services:** A holistic evaluation of the capability of data centres to provide various power system services is missing. This lack of assessment makes it challenging to understand which type of grid services is the best fit to what types of data centres.

To address these gaps, this chapter proposes a method to evaluate the maximum power flexibility of AI-focused HPC data centres and compare it to general-purpose HPC data centres across a wide range of power system services. Additionally, a cost modelling method is proposed for estimating the cost incurred by data centres in providing these services. The cost estimation accounts for the value of computing, utilising data from three major cloud computing platforms: Google Cloud, Amazon AWS, and Oracle. The analysis in this chapter is based on real-world datasets from 7 AI-focused HPC data centres and 7 general-purpose HPC data centres. These datasets span durations of 40 to 80 days, feature temporal resolutions of one second or finer, and include between 13,397 and 962,602 computing jobs. A detailed description of these datasets alongside the access links (all these datasets are publicly available) is provided in Table 6.1.

Using the extensive data above, this chapter finds that AI-focused HPC data centres can provide greater power flexibility and at 50% lower cost than general-purpose HPC data centres for a range of power system services, which enables AI-focused data centres to be more competitive for the same market conditions. This finding is important for both data centre operators and power system operators, and can motivate greater collaboration towards developing more targeted data centre flexibility schemes based on the types of data centres. By comparing flexibility cost and real-world power system service prices, this chapter illustrates the financial profitability of flexibility provision for AI-focused HPC data centres. This may break the stereotype that computing jobs are always more valuable than providing power system services, and brings financial motivations for data centre operators to provide power system services. Additionally, a correlation analysis illustrates that data centre utilisation patterns also

contribute to the greater flexibility and lower cost of AI-focused HPC data centres. Finally, this chapter investigates the opportunity for dynamic quotas for parallelisable jobs to increase data centre flexibility and reduce flexibility provision cost. The superiority of AI-focused HPC data centres still persists in this case.

An advantage of the novel methodology in this chapter is that the flexibility and cost estimates can be scaled using parameters of other data centres through algebraic operations, avoiding the need for re-optimisation. These scaling formulas are detailed in Sections 6.1.1 and 6.1.2. In addition, an interactive Google Colab page [383] is developed to implement this functionality, making the analysis approach in this chapter accessible to anyone to have quick assessment with their own data centre information.

6.1 Methods

This section presents the proposed method to model and evaluate both the maximum flexibility and the corresponding cost for AI-focused and general-purpose data centres. To facilitate understanding, this section is divided into several subsections, each focusing on different case study results presented in this chapter:

1. Section 6.1.1 describes the method for calculating the maximum amounts of flexibility. Related figures include Figs. 6.1c, 6.1d, 6.5, 6.6, and C.1.
2. Section 6.1.2 details the method for calculating the data centre flexibility cost. Related figures include Figs. 6.2, 6.3, 6.4, 6.5, and 6.6.
3. Section 6.1.3 describes the method for calculating the maximum flexibility and the cost with the dynamic quota opportunity (Fig. 6.7).
4. Section 6.1.4 describes the processing of the 14 HPC datasets in Table 6.1.
5. Section 6.1.5 describes the case study settings.

6.1.1 Evaluating the maximum amount of flexibility

This chapter first describes the method for modelling and evaluating the maximum amount of flexibility that data centres can provide for a wide range of power system services.

Flexibility maximisation problem

Power systems require a range of services that DERs such as data centres may provide. Each of these services can be characterised by duration, frequency, and response time [344, 384]. Frequency indicates how often the power system operator activates the service; duration is the length of time that each activation lasts; and response time refers to how quickly a resource delivers the requested flexibility. For example, primary response and tertiary response are two types of power system services that are often used by grid operators to maintain system

frequency within an acceptable range following a demand-supply mismatch. Primary response only requires power delivery to be sustained for a few minutes after activation, but may be activated 15,000 times a year [344]. On the other hand, tertiary response aims to restore the power system to its normal state, and is rarely activated (approximately 20-50 times a year), but when activated may require continuous power delivery for several hours [344]. It should be noted that although electricity markets will generally have similar power system services, their names, delivery requirements, and detailed procurement mechanisms will vary. This chapter focuses only on duration and frequency. Ref. [264] showed that data centres can track power adjustment every few seconds, which is generally sufficient for most power system services.

During the activation periods of each power system service, the data centre flexibility is the difference between the baseline power and the power after adjusting the computing workloads. This workload adjustment can be achieved in practice through job preemption [385,386] and hardware controls for both CPUs and GPUs [291,293,387]. This chapter only considers upward flexibility (i.e., demand reduction), because data centres tend to be highly utilised so downward flexibility potential is very limited. Job completion times could be delayed after adjustment, but a maximum delay limit has been imposed, as was done in [70,250,264].

In this setting, the maximum flexibility for each power system service can be calculated by solving the following optimisation problem:

$$\max_{x_{j,t}, z_{j,t}, n_j^P, p_t, f_t, s_i} \frac{1}{|\mathcal{N}^A|} \sum_{i \in \mathcal{N}^A} s_i \quad (6.1a)$$

s.t.

$$x_{j,t} \in [0, 1], z_{j,t} \in [0, 1], \quad \forall j \in \mathcal{J}, t \in \mathcal{T}, \quad (6.1b)$$

$$x_{j,t} = 0, z_{j,t} = 0, \quad \forall j \in \mathcal{J}, t \notin \mathcal{T}_j, \quad (6.1c)$$

$$z_{j,t} \geq x_{j,t} - x_{j,t+1}, \quad \forall j \in \mathcal{J}, t \in \mathcal{T}, \quad (6.1d)$$

$$n_j^P = \sum_t^T (z_{j,t}) - 1, \quad \forall j \in \mathcal{J}, \quad (6.1e)$$

$$\frac{M^P}{\Delta t} n_j^P \leq \epsilon \cdot D_j, \quad \forall j \in \mathcal{J}, \quad (6.1f)$$

$$\sum_t^T x_{j,t} = D_j, \quad \forall j \in \mathcal{J}, \quad (6.1g)$$

$$\sum_j^J N_j^R \cdot x_{j,t} \leq \hat{N}^R, \quad \forall t \in \mathcal{T}, \quad (6.1h)$$

$$p_t = \sum_j^J G \cdot N_j^R \cdot x_{j,t} + G_0, \quad \forall t \in \mathcal{T}, \quad (6.1i)$$

$$f_t = p_t^{\text{base}} - p_t, \quad \forall t \in \mathcal{T}, \quad (6.1j)$$

$$f_t \geq s_i, \quad \forall t \in \mathcal{T}_i^A, \forall i \in \mathcal{N}^A, \quad (6.1k)$$

$$s_i \geq 0, \quad \forall i \in \mathcal{N}^A \quad (6.1l)$$

The objective (6.1a) is to maximise the amount of the power system service that the data centre can provide on average over the $|\mathcal{N}^A|$ activation events. The constraints are explained below:

(6.1b): This constraint defines the value range for two decision variables. The continuous variable $x_{j,t}$ represents the completed workload of job j at time step t . $x_{j,t} = 1$ represents the maximum workload achievable using the N_j^R user-specified computing resources for a single time step. It is assumed that the computing workload per time step can be continuously adjusted. This is feasible under hardware control techniques such as dynamic power capping or DVFS applicable to both GPUs [291, 292, 387] and CPUs [293, 388]. In addition, real-world data centre schedulers (such as Run:AI) and the NVIDIA MIG functionality support fractional GPU allocation [389, 390]; CPUs are commonly allocated on a fractional basis in current data centres. Therefore, continuity may be better achieved by combining hardware controls and preemption of jobs with fractional resources. The continuous preemption counter variable $z_{j,t}$ is described in the explanation of (6.1d).

Note that, this chapter considers CPUs as the computing resources N_j^R for general-purpose HPC data centres, while GPUs as those for AI-focused HPC data centres due to their importance in AI jobs and higher power consumption [391]. General-purpose HPC data centres also have some GPU jobs, and their CPU utilisation has been considered.

(6.1c): This constraint ensures that both $x_{j,t}$ and the preemption counter variable $z_{j,t}$ are zero outside the available period $\mathcal{T}_j$ of job j . The available period is defined as $\mathcal{T}_j := \{t_j^S, \dots, t_j^C + \lceil (1 + \delta^{\max}) D_j \rceil\}$, determined by the job submission time t_j^S and the completion time t_j^C , incorporating the maximum proportion $\delta^{\max}$ of job delay.

(6.1d): This constraint models the the preemption counter variable $z_{j,t}$. As explained for (6.1b), when the completed workload $x_{j,t+1}$ decreases from the previous timestep $x_{j,t}$, it may be due to: 1) hardware control, and/or 2) preemption of jobs that use fractional resources. In the latter case, extra computing workload is required to save and reload checkpoints. $z_{j,t}$ considers the worst-case scenario and counts each workload decrease as preemption. The boundary condition is $x_{j,T+1} = 0$.

(6.1e): This constraint introduces n_j^P to count the total amount of preemption for each job. The preemption counter $z_{j,t}$ counts the decrease of $x_{j,t}$ for job completion, which is not considered preemption and is thus subtracted from n_j^P .

(6.1f): This constraint ensures that the total extra workloads due to preemption should be minor compared to the computing workload D_j for job completion. Here, ϵ is a small positive number. M^P (minutes) is the extra computing workload per preemption. Δt is the length (minutes) of a single time step.

(6.1g): This constraint ensures job completion. Completing job j requires a computing workload D_j , which is the smallest number of time steps (computing time) required to complete the job using the user-specified number (N_j^R) of computing resources.

(6.1h): This constraint ensures that the total used resources at each time step do not exceed the total number ($\hat{N}^R$) of resources in the data centre.

(6.1i): This constraint establishes a linear relationship between the data centre electric power p_t and the computing workload $x_{j,t}$. The coefficient G in kW represents the increase in power when the utilisation of a single unit of computing resource increases from 0% to 100%, which can include the contribution of memory, cooling, etc. G_0 in kW captures fixed power demand which is independent of computing (e.g. from lighting or idle power). Although certain Google clusters were observed to have non-linear power-utilisation relationships [246], there is also recent work [392] that observed high linear correlation coefficient between GPU power and its utilisation (greater than 0.9 and close to 1) for typical computing workloads. Furthermore, (6.1i)

models the relationship between the total data centre power and the total resource utilisation $\sum_j^J N_j^R \cdot x_{j,t}$, which can be interpreted as the proportion of servers running at the maximum power; in this case, the linear model becomes more reasonable and a demand response controller based on such a model has been demonstrated in a real-world data centre [264].

(6.1j): This constraint models the demand-reduction flexibility f_t . The baseline power p_t^{base} is a parameter and is inferred from the data centre datasets using the same power model (6.1i).

(6.1k): This constrains that the flexibility must be sustained at a specific amount s_i for each service activation indexed by $i \in \mathcal{N}^A$, where the i -th activation spans several consecutive time steps collected in the set $\mathcal{T}_i^A$. The amount of flexibility s_i must be non-negative as specified in (6.1l). Different types of power system service have different frequency $|\mathcal{N}^A|$ and duration $|\mathcal{T}_i^A|$ requirements, where $|\cdot|$ calculates the number of elements in the set. For a specific type of power system service, this chapter sets the duration of each activation to be the same: $|\mathcal{T}_i^A| = |\mathcal{T}_k^A|, \forall i, k \in \mathcal{N}^A$.

Scaling the maximum amounts of flexibility

Problem (6.1) is solved to get the maximum flexibility under a set of nominal data centre parameters. Due to the use of linear power model in (6.1i), for data centres with other parameter settings, one can directly scale the maximum flexibility under the nominal parameters without time-consuming re-optimisation. In this chapter, overhead lines $\overline{}$ are used to specify the nominal parameters. It is worth noting that one must have a specific nominal setting $\overline{G_0} = 0$ for the following derivation. The derived scaling formula can scale the result for arbitrary settings of G_0 .

Assuming all decision variables take the values of the optimal solution, the maximum amount S of flexibility by solving (6.1) can be expressed as:

$$S = \text{ave}_{i \in \mathcal{N}^A} \min_{t \in \mathcal{T}_i^A} f_t. \quad (6.2)$$

The operator $\min_{t \in \mathcal{T}_i^A}(\cdot)$ finds the sustained amount of flexibility during the i^{th} activation. Based on Eqs. (6.1j) and (6.1i), f_t can be expressed as:

$$\begin{aligned} f_t &= p_t^{\text{base}} - p_t \\ &= \sum_j^J G \cdot N_j^R \cdot x_{j,t}^{\text{base}} + G_0 - \sum_j^J G \cdot N_j^R \cdot x_{j,t} - G_0 \\ &= \sum_j^J G \cdot N_j^R \cdot (x_{j,t}^{\text{base}} - x_{j,t}), \end{aligned} \quad (6.3)$$

where the superscript $\blacksquare^{\text{base}}$ denotes values corresponding to the baseline utilisation of the data centre (before flexibility provision). Then, the maximum amount of flexibility can be re-written as:

$$S = \text{ave}_{i \in \mathcal{N}^A} \min_{t \in \mathcal{T}_i^A} \left\{ \sum_j^J G \cdot N_j^R \cdot (x_{j,t}^{\text{base}} - x_{j,t}) \right\}. \quad (6.4)$$

To facilitate comparisons across data centres with different size of power, this chapter specifically focuses on the normalised maximum amounts of flexibility, which is the flexibility in kW divided by the data centre maximum power. Here, the maximum power p^{max} is defined as the power when all computing devices are used 100%. In other words:

$$p^{\text{max}} = G \cdot \hat{N}^R + G_0. \quad (6.5)$$

The normalised flexibility S^{norm} is therefore expressed as:

$$\begin{aligned} S^{\text{norm}} &= \frac{S}{p^{\text{max}}} \\ &= \text{ave}_{i \in \mathcal{N}^A} \min_{t \in \mathcal{T}_i^A} \left\{ \frac{\sum_j^J G \cdot N_j^R \cdot (x_{j,t}^{\text{base}} - x_{j,t})}{G \cdot \hat{N}^R + G_0} \right\}. \end{aligned} \quad (6.6)$$

Then, the normalised flexibility under the nominal parameters (recall that $\overline{G_0} = 0$) can be expressed as:

$$\overline{S^{\text{norm}}} = \text{ave}_{i \in \mathcal{N}^A} \min_{t \in \mathcal{T}_i^A} \left\{ \frac{\sum_j^J \overline{N_j^R} \cdot (x_{j,t}^{\text{base}} - x_{j,t})}{\overline{\hat{N}^R}} \right\}. \quad (6.7)$$

Here this chapter assumes other data centres have similar job characteristics and the only differences are parameters N_j^R , $\hat{N}^R$, G and G_0 . In other words, it is assumed that jobs in other data centres utilise the same proportion of the total data centre resources:

$$\frac{\overline{N_j^R}}{\overline{N_j^R}} = \frac{\overline{\hat{N}^R}}{\overline{\hat{N}^R}}, \forall j \in \mathcal{J} \quad (6.8)$$

Combining Eqs. (6.6), (6.7), and (6.8), the normalised amount of flexibility S^{norm} under arbitrary parameters N_j^R , $\hat{N}^R$, G and G_0 can be calculated by a linear operation of the normalised flexibility under the parameter setting $\overline{S^{\text{norm}}}$:

$$S^{\text{norm}} = \overline{S^{\text{norm}}} \cdot \frac{G \cdot \hat{N}^R}{G \cdot \hat{N}^R + G_0}. \quad (6.9)$$

By combining (6.5), the amount of flexibility S in kW after scaling is:

$$S = S^{\text{norm}} \times p^{\text{max}} = \overline{S^{\text{norm}}} \cdot G \cdot \hat{N}^R. \quad (6.10)$$

Eq. (6.9) suggests that the normalised maximum amounts of flexibility results are independent of data centre parameters, as long as there is negligible fixed power that is independent of computing (i.e., $G_0 = 0$). This observation supports the use of normalised maximum flexibility as the evaluation metric in this chapter, as defined in Eq. (6.6).

It is important to note that, while the assumption stated in (6.8) could produce imprecise scaling results if other data centres exhibit different job characteristics, the scaling formula can still help other data centres to quickly evaluate their flexibility by scaling the results calculated in the nominal setting (or via the interactive Google Colab page [383] developed in this thesis).

6.1.2 Data centre cost of flexibility

This section describes the proposed method for modelling the cost of data centre flexibility, along with the approach for estimating this cost using real-world datasets.

Cost modeling

This thesis assumes that data centres charge a price for each computing job, as is the case with cloud computing platforms. This thesis further assumes that the price will be lower if there is greater delay in job completion, due to reduced timeliness. Thus, the total cost for the data centre providing a particular power system service is the sum of the price reduction incurred by the delays across all jobs given optimal job scheduling. Note that the proposed cost model is not limited to cloud computing data centres, because other data centres may also implement pricing schemes for internal usage; delays in computing-related projects (e.g., developing systems such as ChatGPT) could similarly be monetised with appropriate financial data and analysis—however, these data are typically only available for cloud computing data centres. Nevertheless, it is expected that any deviations from the cost estimates in this chapter to be relatively small for the same computing hardware, because such hardware (e.g., GPUs or CPUs) would otherwise not be rentable through cloud computing data centres.

Based on the modelling above, this thesis proposes using the following linear model to express the price reduction c_j for each job j :

$$c_j = A_j \cdot \delta_j \cdot V_j \quad (6.11)$$

where V_j is the original computing price of the job, A_j is the price reduction coefficient, measuring the proportionate price reduction in response to a certain delay proportion δ_j . The delay proportion δ_j is defined as the ratio of the extra time t_j^{Ex} beyond the original job completion time relative to the computing time D_j :

$$\delta_j = \frac{t_j^{Ex}}{D_j} \quad (6.12)$$

In several cloud computing platforms such as Google Cloud, AWS, Oracle, and Lambda Cloud, the computing price V_j is calculated as:

$$V_j = D_j \cdot R \cdot N_j^R \quad (6.13)$$

where R is the hourly price of a single unit of computing resource and N_j^R is the user-specified number of resources. Recall that the computing resources refer to GPUs for AI-focused HPC data centres and refer to CPUs for general-purpose HPC data centres. Also note that most cloud GPU prices will include necessary CPU, memory, and storage.

As explained above, the total flexibility cost C^{Flex} is:

$$C^{\text{Flex}} = \sum_{j \in \mathcal{J}} c_j \quad (6.14)$$

Cost minimisation problem

This thesis assumes that data centre operators are motivated to provide a certain power system services at the minimum cost. Therefore, the following problem is solved to find the data centre cost of providing S amount of flexibility for a specific power system service.

$$\min_{x_{j,t}, x'_{j,t}, z_{j,t}, n_j^P, p_t, f_t, s_i, e_j, \delta_j, c_j} C^{\text{Flex}} \quad (6.15a)$$

s.t.

$$(6.1b) - (6.1l), \quad (6.15b)$$

$$x'_{j,t} \geq x_{j,t}, \quad \forall j \in \mathcal{J}, t \in \mathcal{T}, \quad (6.15c)$$

$$e_j \geq t \cdot x'_{j,t} + 1, \quad \forall t \geq t_j^S + D_j, \forall j \in \mathcal{J}, \quad (6.15d)$$

$$\delta_j \geq 0, \delta_j \geq \frac{e_j - t_j^S - D_j}{D_j}, \quad \forall j \in \mathcal{J}, \quad (6.15e)$$

$$c_j \geq A_j \cdot \delta_j \cdot D_j \cdot R \cdot N_j^R, \quad \forall j \in \mathcal{J}, \quad (6.15f)$$

$$\frac{1}{|\mathcal{N}^A|} \sum_{i \in \mathcal{N}^A} s_i \geq S \quad (6.15g)$$

Constraint (6.15c) introduces binary variables $x'_{j,t}$ to indicate the job running status. In (6.15d), the decision variable e_j models the last time step of job running. Therefore, the term $e_j - t_j^S - D_j$ in (6.15e) represents the time delay t^{Ex} . Constraint (6.15f) models the reduction in computing price as defined in (6.11). Constraint (6.15g) ensures that the average flexibility over the $|\mathcal{N}^A|$ times of service activation can meet the flexibility requirement S .

Note that cost minimisation problem (6.15) is a challenging mixed-integer linear programming (MILP), and the solution process can be speeded up based on the proposed tightening trick detailed in Appendix C.1.

The average cost of data centre flexibility

Solving problem (6.15) leads to the total flexibility cost C^{Flex} . Based on C^{Flex} , this chapter further calculates the *average flexibility cost* ($ACoF$), which is the total flexibility cost divided by the total shifted energy during the service activation periods in kWh (the product of the flexibility amount, duration, and frequency). If the average flexibility cost is less than the price of the power system service, then providing that service is profitable. Therefore, the average flexibility cost $ACoF$ provides more intuitive interpretation than the total flexibility cost C^{Flex} , and will be evaluated in all the cost plots of this chapter.

Scaling the cost of flexibility provision

Problem (6.15) is solved to get the minimum flexibility cost C^{Flex} , which then leads to the average flexibility cost $ACoF$ under a set of nominal data centre parameters. Due to the use of linear cost model, for other data centre parameters, one can directly scale the $ACoF$ calculated under the nominal parameters without re-optimisation. Similar to Section 6.1.1, overhead lines ■ are used to specify the nominal parameters.

Again, this section assumes that all decision variables take the values of the optimal solution. For subsequent derivation, this chapter assumes that all jobs have the same price reduction coefficient $A = A_j$. In practice, the value of A can be calculated as a weighted average of the coefficients of A_j for all jobs. Then based on Eq. (6.11), the (minimum) total cost of providing a certain amount S of flexibility can be expressed as:

$$C^{\text{Flex}} = \sum_{j \in \mathcal{J}} c_j = A \cdot \sum_{j \in \mathcal{J}} \delta_j \cdot D_j \cdot R \cdot N_j^{\text{R}} \quad (6.16)$$

The total shifted energy E^f in kWh over the service activation periods can be expressed as:

$$E^f := \Delta t \cdot |\mathcal{N}^A| \cdot |\mathcal{T}_i^A| \cdot S \quad (6.17)$$

Based on Eq. (6.10), it follows that:

$$E^f = \Delta t \cdot |\mathcal{N}^A| \cdot |\mathcal{T}_i^A| \cdot \overline{S^{\text{norm}}} \cdot \hat{N}^{\text{R}} \cdot G \quad (6.18)$$

Note that, due to the use of (6.10), one needs to set the nominal parameter $\overline{G_0} = 0$ to maintain the initial assumption. Given (6.16) and (6.18), it follows that:

$$ACoF = \frac{C^{\text{Flex}}}{E^f} = \frac{A \cdot \sum_{j \in \mathcal{J}} \delta_j \cdot D_j \cdot R \cdot N_j^{\text{R}}}{\Delta t \cdot |\mathcal{N}^A| \cdot |\mathcal{T}_i^A| \cdot \overline{S^{\text{norm}}} \cdot \hat{N}^{\text{R}} \cdot G} \quad (6.19)$$

Again, if (6.8) holds, the average flexibility cost under the nominal parameter setting can be expressed as:

$$\overline{ACoF} = \frac{\bar{A} \cdot \sum_{j \in \mathcal{J}} \delta_j \cdot D_j \cdot \bar{R} \cdot N_j^R}{\Delta t \cdot |\mathcal{N}^A| \cdot |\mathcal{T}_i^A| \cdot \bar{S}^{\text{norm}} \cdot \hat{N}^R \cdot \bar{G}} \quad (6.20)$$

Combining (6.19) and (6.20), given $\overline{ACoF}$ calculated in the nominal setting, the formula to scale the $ACoF$ according to other data centre parameters A , R , and G is:

$$ACoF = \overline{ACoF} \cdot (A \cdot R \cdot \bar{G}) / (G \cdot \bar{A} \cdot \bar{R}) \quad (6.21)$$

The term $(A \cdot R \cdot \bar{G}) / (G \cdot \bar{A} \cdot \bar{R})$ is the *cost scaling factor*, which can be used to scale up the calculated $\overline{ACoF}$ in a nominal setting based on other data centre parameters.

Estimating the cost scaling factor

The sections above provide the method for calculating the flexibility cost in a nominal setting (can be set freely with $\bar{G}_0 = 0$), and describe how to scale the cost to other parameter settings with the cost scaling factor. However, a question remains on how to estimate a cost scaling factor that allows to adjust the nominal cost result to reflect the flexibility costs of real-world data centres. Answering this question is the foundation of all the cost plots, such as Figs. 6.2, 6.3, 6.4, 6.5, and 6.6 that will be introduced in subsequent content. This section describes the estimation method based on real-world cloud platform data.

As discussed in the early part of this section, the flexibility cost is based on the assumption that slower computing will lead to a lower computing price. This assumption aligns to the fact that cloud platforms provide several CPU and GPU rental options with different computing speeds and prices; the cost scaling factor can then be estimated based on these information.

The estimation method is illustrated by the following example. Suppose a cloud platform provides two computing rental options. Option I has N_1^R computing resources, a computing speed of O_1 , hourly price per computing resource of R_1 , and the power per computing resource is G_1 (in kW), while the parameters for Option II are N_2^R , O_2 , R_2 , and G_2 . Suppose a computing job has computing workload of W . Then, if Option I is rented, the computing time is $D_1 = W/O_1$. Based on (6.13), the computing price is $V_1 = W/O_1 \cdot R_1 \cdot N_1^R$. Similarly, if Option II is rented, the computing time is $D_2 = W/O_2$, and the price is $V_2 = W/O_2 \cdot R_2 \cdot N_2^R$. Without loss of generality, we assume that Option I is the faster and more expensive option, namely $D_1 < D_2$ and $V_1 > V_2$. Now, if Option I is used to provide power system flexibility service and the computing time is increased to $D'_1 = D_2$ (job delay), then the computing price of the delayed Option I should not exceed the price of Option II. In other words, the flexibility cost of Option I is:

$$c_1 = V_1 - V_2 \quad (6.22)$$

By definition, the delay proportion of Option I for providing flexibility is $\delta_1 = (D_2 - D_1)/D_1$. Then, based on Eq. (6.11), the price reduction coefficient of Option I can be estimated as:

$$A_1 = c_1/(\delta_1 \cdot V_1) \quad (6.23)$$

Then, based on (6.21), the corresponding cost scaling factor can be estimated as $(A_1 \cdot R_1 \cdot \bar{G})/(G_1 \cdot \bar{A} \cdot \bar{R})$.

Through the example above, it can be seen that the cost scaling factor can be obtained by comparing two computing rental options. Therefore, this chapter collected all GPU and CPU rental options available in Google Cloud, AWS, Oracle, and Lambda Cloud. It should be noted that cloud platforms provide CPUs not just for HPC but also for other applications such as web services. As this chapter focuses on HPC data centres, only a subset of CPU options on cloud platforms with optimised computing performance is selected. Details of the collection process can be found in Appendix C.4. The CPU speed information can be found on the PassMark website [393]. The CPU mark is used as the speed information, which is the weighted harmonic average of the computing speed for several benchmark tests (e.g., integer math, floating point math, physics) that are important for HPC applications [394]. The GPU speed information is available at Lambda Lab [395], which is the weighted average speed for a range of AI tasks. The speed information is provided on both the FP32 and FP16 basis, and the harmonic average of the two is used as the final GPU speed. More than 100 CPU/GPU options with the complete desired information are recorded. This chapter sets G as the device rated power. Note that, as the utilisation of a device (GPU or CPU) increases from 0 to 100%, the power consumption may be lower than the device rated power due to the idle power, leading to an overestimation of G . However, there are also power increases in memory and cooling that may counterbalance the overestimation of G to some extent. Finally, on most cloud platforms, CPUs are rented on the basis of virtual CPUs (vCPUs), which typically represent one thread of a physical CPU. The power of a vCPU is calculated by dividing the rated power of the physical CPU by the number of vCPUs.

With more than 100 CPU/GPU options from these cloud service providers, for either CPU (represents general-purpose data centres) or GPU (represents AI-focused data centres), this chapter compares any two options within the same provider to estimate the corresponding cost scaling factor, as illustrated in the previous example. This chapter avoids comparing options across different platforms because their additional services also affect the pricing. This chapter does not compare the computing options if one would require 100% more computing time than the other. This is to exclude options that are excessively slow and therefore less comparable. In addition, there are comparisons where one option is slower and more expensive than the other, and these comparisons are dropped. These “slower and more expensive”

options exist because (potentially) there is a shortage in other computing resources or they have larger memory or storage spaces. Overall, in this chapter, 38 samples of the cost scaling factor were collected for general-purpose HPC data centres, and 55 samples were for AI-focused HPC data centres.

Note that, as mentioned in the main content, the Lambda Cloud GPU prices are significantly lower than the other three cloud platforms. To minimise potential bias, as will be demonstrated in Fig. 6.4, the case study will separate the analysis of Lambda Cloud data from the other three cloud platforms.

6.1.3 Dynamic quota

Certain computing jobs, such as AI training with batch computation, are parallelisable. Therefore, as long as data centre computing resources are not always 100% utilised, they can leverage unused resources to speed up computations, reducing job delays and thus flexibility cost. This is called “dynamic quota” as in [396] and is used by data centre schedulers [386]. This chapter takes this additional opportunity into account, and provides a method to model this opportunity in this section.

Solving the maximum flexibility under dynamic quota

To model the dynamic quota functionality, this chapter introduces another set of continuous variables $x_{j,t}^{\text{DQ}} \in [0, 1]$ and a parameter $K^{\text{DQ}} \in [0, 1]$ that represents the speed-up coefficient. In other words, using 100% more computing resources ($x_{j,t}^{\text{DQ}} = 1$) will complete K^{DQ} more computing workload in a single time step. The problems to find the maximum amount of flexibility with dynamic quota have the following set of updated constraints:

- (6.1g) is updated to

$$\sum_t^T (x_{j,t} + K^{\text{DQ}} x_{j,t}^{\text{DQ}}) = D_j \quad (6.24)$$

- (6.1h) is updated to

$$\sum_j^J N_j^{\text{R}} \cdot (x_{j,t} + x_{j,t}^{\text{DQ}}) \leq \hat{N}^{\text{R}} \quad (6.25)$$

- (6.1i) is updated to

$$p_t = \sum_j^J G \cdot N_j^{\text{R}} \cdot (x_{j,t} + x_{j,t}^{\text{DQ}}) + G_0 \quad (6.26)$$

- Adding more computing resources may not always result in speed-up due to communication delays across devices. To avoid overestimating the capability of dynamic quota, the number of additional resources is limited to not exceed the originally specified resources through:

$$x_{j,t}^{\text{DQ}} \leq x_{j,t} \quad (6.27)$$

Calculating the flexibility cost under dynamic quota

As extra computing resources typically do not achieve a full 100% speed-up, activating the dynamic quota would result in additional energy costs, alongside the flexibility cost triggered by the reduction in computing prices as defined in (6.14). The extra energy cost can be expressed as:

$$C^E = \hat{C}_{\text{elec}} \Delta t \left(\sum_{t \in \mathcal{T}} p_t - \sum_{t \in \mathcal{T}} p_t^{\text{base}} \right) \quad (6.28)$$

where $\hat{C}_{\text{elec}}$ is the energy price and $\Delta t (\sum_{t \in \mathcal{T}} p_t - \sum_{t \in \mathcal{T}} p_t^{\text{base}})$ is the additional energy consumption due to the dynamic quota. The total flexibility cost with the dynamic quota opportunity becomes:

$$C^{\text{Flex}} = \sum_{j \in \mathcal{J}} c_j + C^E \quad (6.29)$$

Eqs. (6.28) and (6.29) are included in the MILP-based cost minimisation problem (6.15) to calculate the flexibility cost under dynamic quota. Similar to Section 6.1.2, this thesis also introduce a tightening trick for the MILP under dynamic quota, as detailed in Appendix C.2.

Scaling the cost under dynamic quota

With the dynamic quota opportunity, there will be extra energy cost, so the *ACoF* would additionally include the average extra energy cost of flexibility (*AECof*). To differentiate, this chapters refer to the *ACoF* without dynamic quota (Eq. (6.21)) as *APCoF*. The *ACoF* with the dynamic quota opportunity is the sum of *APCoF* and *AECof*. The scaling of *APCoF* still follows (6.21). As for *AECof*, first, based on Eq. (6.1i), the extra energy cost C^E in (6.28) can be re-express as:

$$C^E = \hat{C}_{\text{elec}} \Delta t \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} G \cdot N_j^R \cdot (x_{j,t}^{\text{base}} - x_{j,t}) \quad (6.30)$$

Combined with Eq. (6.18), *AECof* can be expressed as:

$$AECof = \frac{C^E}{Ef} = \frac{\hat{C}_{\text{elec}} \Delta t \sum_{t \in \mathcal{T}} \sum_j^J \cdot N_j^R \cdot (x_{j,t}^{\text{base}} - x_{j,t})}{\Delta t \cdot |\mathcal{N}^A| \cdot |\mathcal{T}_i^A| \cdot \bar{S}^{\text{norm}} \cdot \hat{N}^R} \quad (6.31)$$

Again this chapter assumes (6.8) holds, following a similar process, given the $\overline{AECof}$ calculated in the nominal parameter $\bar{\hat{C}}_{\text{elec}}$, the formula to scale to other parameter settings is:

$$AECof = \overline{AECof} \cdot \frac{\hat{C}_{\text{elec}}}{\bar{\hat{C}}_{\text{elec}}} \quad (6.32)$$

Therefore, the scaling formula for $ACoF$ with the dynamic quota opportunity is:

$$\begin{aligned} ACoF &= APCoF + AECof \\ &= \overline{APCoF} \cdot \frac{A \cdot R \cdot \overline{G}}{G \cdot \overline{A} \cdot R} + \overline{AECof} \cdot \frac{\hat{C}_{elec}}{\overline{C}_{elec}} \end{aligned} \quad (6.33)$$

6.1.4 Data processing

This section presents data processing of all 14 HPC data centre datasets used in the case studies of this chapter. Details of these 14 data centres are listed in Table 6.1. Some background information is provided. There are three important time steps for each job record: the submission time, start time, and completion time. The submission time is when the user commits the job. The start time is when the submitted job begins running, which may be later than the submission time due to resource unavailability. The completion time is when the job is completed. The difference between the completion time and the start time is the job computing time. Because some of the data centre datasets lack job submission times, for consistent comparisons, this chapter sets all submission times equal to the start times, resulting in zero queue time. This can further lead to an underestimation of the data centre flexibility, since there could have been queue-time flexibility.

Data selection

Among the 14 HPC data centre datasets, some of them contain job data that span several years. To reduce simulation complexity and ensure up-to-dateness, this chapter selected 80 consecutive days of job records from the most recent year. Note that, the data centre recorder might miss jobs that were submitted before or completed after the recording period, which leads to under-utilisation at the beginning or the end of the recording period. These under-utilised periods are trimmed accordingly. Some datasets contain fewer than 80 days of records post-trimming, thus only 40 or 60 consecutive days are chosen.

Short-duration jobs

Job start times and completion times are rounded to the time resolution Δt . For jobs having zero computing time after rounding, this chapter sets their computing times to one time step and rescale the required number of resources to ensure the same computing workload: the product of the computing time and the required number of resources.

Job partition

For jobs started before and/or completed after the optimisation horizon $\mathcal{T}$, this chapter truncates their computing times to equal the number of time steps of job running within $\mathcal{T}$ under the baseline profiles.

For example, suppose the optimisation horizon is 00:00 Day 1 - 24:00 Day 1, if there is a job started at 18:00 Day 0 (6 hours before the horizon) and is supposed to complete at 2:00 Day 2 (2 hours after the horizon), then the computing time of this job is set to 24 hours.

As the second example, considering the same optimisation horizon, if there is a job started at 18:00 Day 1 and is supposed to complete at 2:00 Day 2 (2 hours after the horizon), in this case, this job only needs computation for 24:00D1 - 18:00D1 = 6 hours.

Job aggregation

Some of the 14 HPC datasets contain numerous jobs that are running simultaneously, which leads to complicated optimisation problems used to find the maximum amount of flexibility (6.1) and the minimum flexibility cost (6.15). Since this paper focuses on high-level evaluation rather than production-level implementation, this chapter uses a job aggregation strategy to reduce the computational complexity while maintaining the computing pattern.

This chapter aggregates daily jobs into 100 groups using K-means clustering based on the start time and completion time, then replaces each group with a single aggregated job, resulting in only 100 jobs per day after aggregation. For the aggregated job of each group, the start time and the submission time are the earliest start and submission times of the original jobs in the group, and the completion time is the latest. The computing time of the aggregated job is the difference between the completion and start times. The number of resources (GPUs/CPU) used for each aggregated job is calculated by summing the products of computing time and the number of resource for each job j in the group, then dividing by the computing time of the aggregated job. This process ensures that the computing workload of the aggregated job is equal to the total workloads of the original jobs in that group.

Note that the use of k-means is to reduce the job details to a certain resolution level so as to speed up computation. The data centre job scheduling results based on the aggregated jobs may not satisfy constraints for the original jobs. To see this, we can view each job as a battery (the job completion constraint mimic the energy requirement constraint), and existing DER aggregation papers [29, 95] have demonstrated that summing over battery parameters lead to an unsafe outer approximation (also see Chapter 4). With this in mind, the consistent setting of 100 groups for all data centres ensures that all data centres are under the same resolution of job details, which in turn ensure a like-for-like comparison. Due to this reason, the number of groups is not determined by methods for “the optimal number of cluster” in standard k-means applications, such as by the elbow curves. In fact, “the optimal number of cluster”

methods can be less applicable here. For example, consider a data centre dataset where all jobs have similar start times and completion times and “the optimal number of cluster” is 1; this aggregates all jobs of a data centre into a single one, neglecting all individual job constraints and lead to considerably unsafe outer approximation.

The aggregation strategy improves computational efficiency while preserving utilisation fidelity. Appendix C.4 plots the utilisation curves based on the aggregated jobs and those based on the original jobs.

6.1.5 Problem settings

The optimisation horizon is set to 10 days with a resolution Δt of 15 minutes, resulting in $T = 960$. In the 10-day optimisation horizon, $|\mathcal{N}^A|$ non-overlapping flexibility activation periods are randomly picked from a uniform distribution, where each period contains $|\mathcal{T}_i^A|$ consecutive time steps. This achieves an annual flexibility frequency of $365/10 \times |\mathcal{N}^A|$ with a duration of $|\mathcal{T}_i^A|$ time steps per activation. This chapter repeats the 10-day optimisation for the whole time span of the dataset, and the final results are the average based on these runs.

The overhead M^P in (6.1f) per preemption is 1.5 minutes for AI-focused HPC data centres based on [283], and 0.5 minutes for general-purpose HPC data centres, considering the avoidance of GPU communication compared to AI-focused HPC data centres. The extra workload due to preemption is limited below 1% of the original computing workload, namely $\epsilon = 1\%$ in (6.1f).

The nominal parameters are set to $\overline{G} = 1$ kW, $\overline{G}_0 = 0$, $\overline{A} = 0.5$, $\overline{R} = 1$ USD per hour per computing resource. The energy price $\widehat{C}_{\text{elec}}$ is set to the wholesale market price (front-of-the-meter) of 0.05 USD/kWh [344], considering the excessive power demand of hyperscale data centres.

6.1.6 Summary of assumptions

For the convenience of reading, this section summarises the key assumptions made in this chapter and discusses their potential impacts. Detailed reasons and justifications for these assumptions can be found in the corresponding sections of this chapter.

1. Constraint (6.1i): This thesis assumes a linear relationship between data centre computing power and computing resource utilisation. The linear power model is supported by recent work [392], which observed a high linear correlation coefficient (greater than 0.9 and close to 1) between GPU power and its utilisation under typical computing workloads. However, this relationship may depend on specific data centre hardware. For example, some Google data centres have been observed to exhibit non-linear power-utilisation relationships [246].

2. Constraint (6.1k): When calculating the maximum flexibility for each power system service, this thesis assumes that the service is 100% activated throughout the service period. In practice, power system services such as reserves are typically only partially activated. Therefore, this assumption may underestimate the maximum data centre flexibility.
3. Scaling assumptions (Sections 6.1.1 and 6.1.2): The validity of the flexibility and cost scaling formulas relies on the assumption that jobs in other data centres utilise the same proportion of total data centre resources, as expressed in Eq. (6.8). These formulas are intended solely for the convenient transfer of our results to other data centres, and do not affect the main findings of this chapter.
4. Section 6.1.2: The thesis assumes that slower computing results in a lower computing price for the same computing job. This forms the basis of the proposed data centre cost model. The assumption is consistent with the observation that faster GPUs or CPUs are typically priced higher (assuming similar specifications such as memory size and precision) in current markets and cloud platforms.
5. Section 6.1.4: Due to missing job submission times in some datasets, this chapter sets all submission times equal to job start times, ensuring consistent comparisons. This results in zero queue time, which may lead to an underestimation of data centre flexibility, as potential queue-time flexibility is not captured.

6.2 Results

In this section, the maximum flexibility and the associated cost of data centre flexibility are first analysed for a wide range of power system services, using one typical AI-focused data centre and one typical general-purpose data centre as case studies. An extended analysis based on 14 data centres is then conducted to generalise the findings, after which the impact of distinct job patterns on data centre flexibility and cost are analysed. Additionally, this chapter analyses the dynamic quota scenario where extra computing resources can be activated to speedup computation, demonstrating its potential to enhance data centre flexibility.

6.2.1 Maximum amount of data centre flexibility

Due to the computational complexity of solving the optimisation problems, this chapter starts by comparing the results for one general-purpose HPC data centre (ORNL in Table 6.1) and one AI-focused HPC data centre (Saturn in Table 6.1). These two data centres have average utilisation rates both at 81%, and Figs. 6.1a and 6.1b shows their baseline utilisation time series.

Table 6.1: Summary of the analysed HPC data centres

Name	Type	Main applications	Timespan	Source	System specs	Type
Seren	AI	LLM (training/development), 7B to over 123B [391]	03/2023 to 08/2023	Shanghai AI Lab [397]	8*286 GPUs (A100), 128*286vCPUs, 1024*286 GB RAM	Academic
Philly	AI	NN training only, CNN, RNN, LSTM [398]	08/2017 to 12/2017	Microsoft's internal Philly clusters [399]	2490 GPUs	Commercial
PAI	AI	DL, RL, training and inference [400]	07/2020 to 08/2020	Alibaba [401]	6742 GPUs (T4/P100/V100), 156576vCPUs	Commercial
Saturn	AI				2096 GPUs (Pascal & Volta), 64*262vCPUs, 256*262GB RAM	
Earth	AI	DL for CV, NLP, and RL. Various types of jobs in the DL development pipeline, e.g., data preprocessing, model training, inference, quantisation, etc, but the majority is for training [71].	04/2020 to 09/2020	SenseTime [402]	1144 GPUs (V100), 48*143vCPUs, 376*143GB RAM,	Commercial
Venus	AI				1064 GPUs (V100), 48*133vCPUs, 376*133GB RAM,	
Uranus	AI				2112 GPUs (Pascal), 64*264vCPUs, 256*264GB RAM	
ORNL	General	General scientific computing, like Fluid Dynamics and physics accelerators (recorded in their data files) [403]	01/2015 to 12/2019	ORNL [403]	18688 nodes, 299,008 CPU cores, 18688 Kepler K20X GPUs [404]	Academic
MIRA	General	General scientific computing, scientific research, including studies in the fields of material science, climatology, seismology, and computational chemistry [405].	04/2013 to 03/2020	Argonne Leadership Computing Facility [406]	786432 CPU cores	Academic
THETA	General	Simulation, data analytics, AI [407], detailed molecular simulations to massive cosmological models [408]	07/2017 to 01/2024		281088 CPU cores for THETA + 24*128 (EPYC 7742) for THETA GPU	
CIEMAT-Euler	General	Data science, AI applied to medicine, dark matter, proton collider analysis and Virgo gravitational waves identification, modelling and simulating of fluid mechanics: combustion, flame dynamics, acoustics [409]	11/2008 to 12/2017	Parallel Workloads Archive [410]	1920 CPU cores (may include GPUs as the application includes AI, but not clearly stated)	Academic
PM100	General	Computational Chemistry, Condensed Matter Physics and Computational Fluid Dynamics [411], advanced fluid-dynamic simulations [412], Nuclear Physics [413], drug discovery [414].	05/2020 to 10/2020	Zenodo [415]	980 nodes, 32 CPU cores, 4 V100 GPUs	Academic
University of Luxembourg Gaia Cluster	General	It is used mainly by biologists working with large data problems and engineering people working with physical simulations [410].	05/2014 to 08/2014	Parallel Workloads Archive [410]	2004 CPU cores	Academic
Meta-centrum 2	General	Computer science, computational chemistry, biomedical computing, bioinformatics, physics, bioinformatics, computer science (middleware technologies), interoperability (research infrastructures) computer science (data platforms) [416].	01/2013 to 12/2014	Parallel Workloads Archive [410]	19 clusters, with 495 nodes and 8412 cores (include some GPU nodes)	Academic

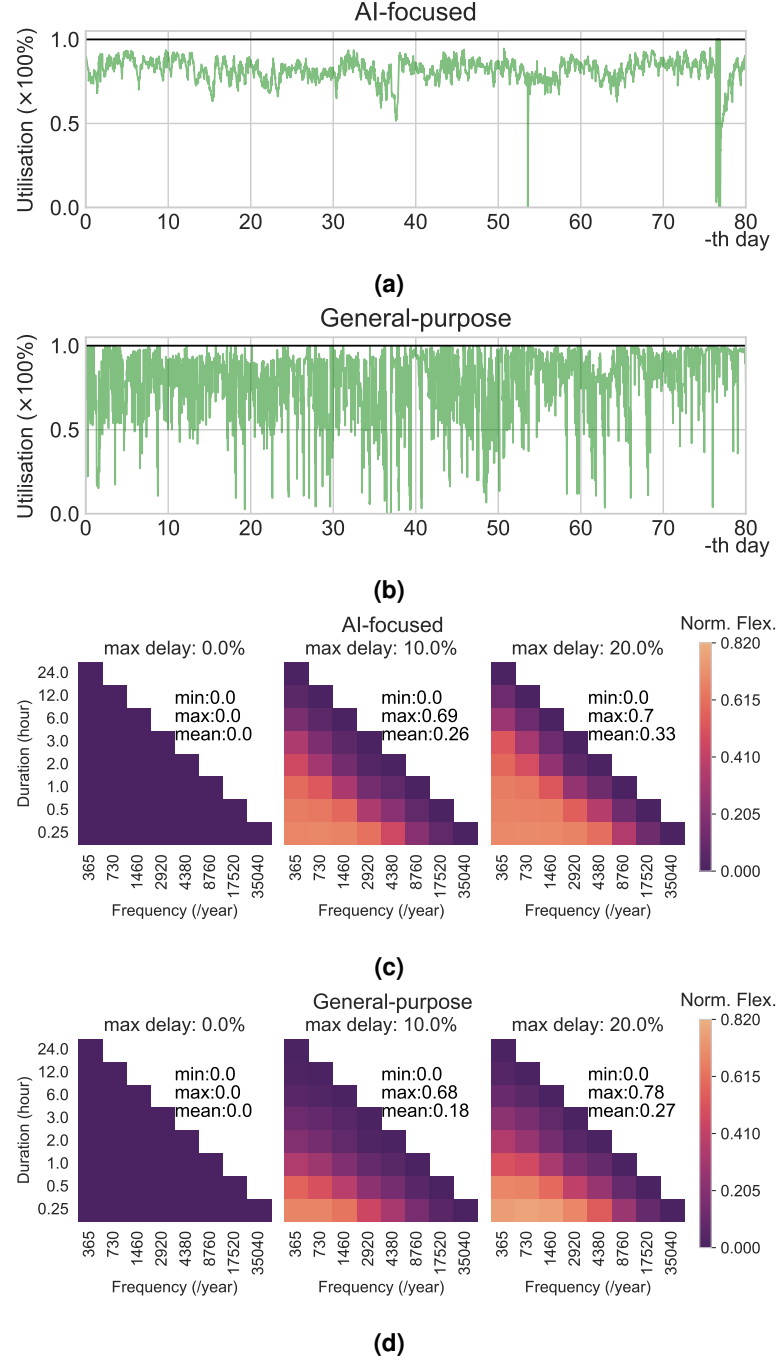


Figure 6.1: Baseline utilisation time series and maximum amount of flexibility for two selected data centres.

(a) Utilisation of the AI-focused HPC data centre (Saturn in Table 6.1). (b) Utilisation of the general-purpose HPC data centre (ORNL in Table 6.1). The black lines represent the utilisation upper bound, while the green lines show the utilisation time series.

(c) and (d) are for the maximum amounts of flexibility for power system services with different duration and frequency requirements. (c) Results for the AI-focused HPC data centre (Saturn). (d) Results for the general-purpose HPC data centre (ORNL). The maximum flexibility amounts are normalised ('Norm. Flex') as ratios of the maximum power of the data centre, following Eq. (6.6). The heat maps in each column are subject to a specific maximum delay limit (max delay), which is the maximum delay time proportional to the job computing time. "min", "max", and "mean" refer to the minimum, maximum, and mean values across all blocks in each heat map.

Fig. 6.1c shows the maximum amounts of flexibility calculated for power system services with different frequency and duration requirements for the AI-focused HPC data centre (Saturn), and 6.1d shows the results for the general-purpose HPC data centre (ORNL). These flexibility amounts are normalised by the maximum data centre power, as per Eq. (6.6). In this chapter, the time resolution of optimisation problems is set to 15 minutes for computational tractability. Therefore, Figs. 6.1c and 6.1d have considered all possible duration and frequency requirements in this setting. These requirements also cover all the 12 typical power system services listed in Ref. [344]. Comparing Figs. 6.1c and 6.1d, it is apparent that the AI-focused HPC data centre provides greater flexibility than the general-purpose HPC data centre as the service duration increases above 1 hour. This is because the general-purpose HPC data centre has more variable baseline utilisation (Figs. 6.1a and 6.1b), so it is more difficult to sustain demand reduction flexibility for a long time. This difference in utilisation patterns is also observed in the analysis for all 14 data centres, as will be illustrated in Fig. 6.6. Note that, Figs. 6.1c and 6.1d show that the general-purpose HPC data centre has slightly greater flexibility for short-duration services. In fact, as discussed in Section 6.1.4, this is because the job partition strategy in this chapter underestimates the flexibility of the AI-focused HPC data centre with more long-computing-time jobs. As demonstrated in Appendix C.3, these two data centres have similar amounts of flexibility for short-duration services with a sufficiently long optimisation horizon.

6.2.2 Cost of data centre flexibility

Figs. 6.2a–6.2f show the average flexibility cost $ACoF$ for various power system services when data centres provide different percentages of their maximum flexibility (as evaluated in Figs. 6.1c and 6.1d). These figures also display results under different percentiles of the cost scaling factor derived from the samples. It can be seen that the AI-focused HPC data centre has lower flexibility cost than the general-purpose HPC data centre for all power system services and all percentiles of the cost scaling factor. Specifically, when the cost scaling factor is at the median (P50), the AI-focused HPC data centre shows a lower flexibility cost of at least 50%. One reason for the lower flexibility cost is the smaller cost scaling factor of AI-focused HPC data centres. Fig. 6.3 shows the distribution of the cost scaling factor and its components. One interesting observation is that, although GPUs (representing AI-focused HPC data centres) are more expensive than CPUs (representing general-purpose HPC data centres), the GPU power consumption is also large, which leads to a lower price-to-power ratio than CPUs (the R/G plot), which in turn brings a lower cost scaling factor to AI-focused HPC data centres. This observation is also consistent with the R and G plots in Fig. 6.3: when comparing the R and G plots, CPUs maintain a similar relative position along the y-axis, while GPUs exhibit an upward y-axis shift in the G plot, reflecting their greater computational power. Consequently, GPUs display a lower R/G ratio in the corresponding R/G plot.

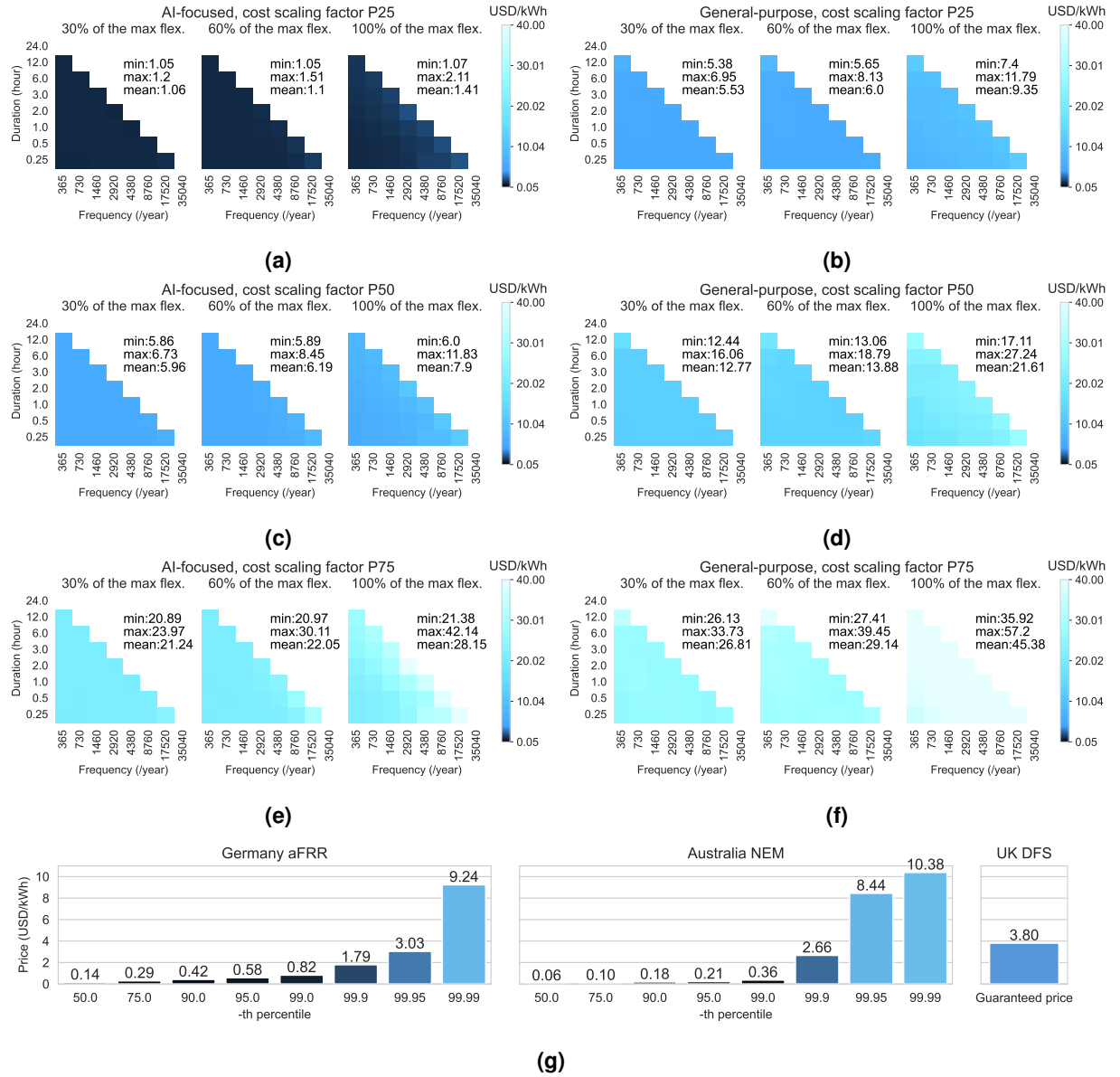


Figure 6.2: Average flexibility cost and power system service prices.

In (a)-(f), each heatmap displays the average cost of providing different percentages of the maximum amounts of flexibility evaluated in Figs. 6.1c and 6.1d. The maximum delay limit is set to 20%. “min”, “max”, and “mean” refer to the minimum, maximum, and mean values across all blocks in each heatmap. Plots (a), (c), and (e) correspond to the AI-focused HPC data centre (Saturn) under 25-th (P25), 50-th (P50), and 75-th (P75) percentiles of the cost scaling factor, estimated using data from Google Cloud, AWS, and Oracle. Plots (b), (d), and (f) correspond to the general-purpose HPC data centre (ORNL). Plot (g) shows percentiles of real-world power system service prices from Germany’s aFRR [417], Australia’s NEM [418], and UK’s DFS [419].

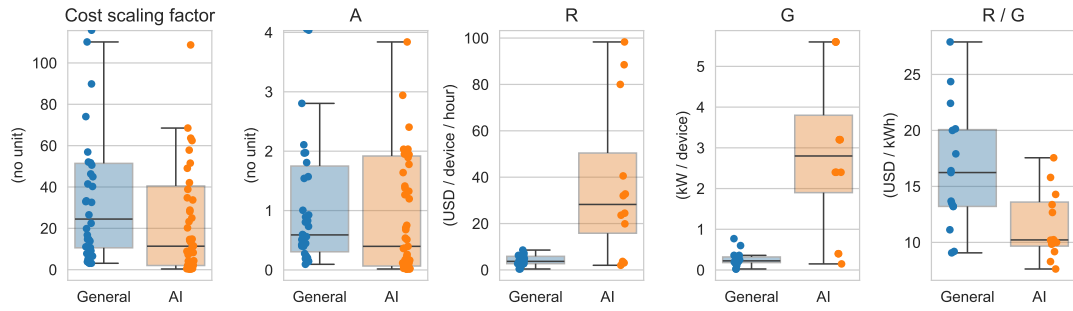


Figure 6.3: Samples of the cost scaling factor and the elements (A , R , and G) involved in its computation (see Eq. (6.21)).

These samples are derived based on data of computing rental options on Google Cloud, AWS, and Oracle. The 25th, 50th, and 75th percentiles of the cost scaling factor for general-purpose HPC data centres (General) are 10.60, 24.49, and 51.42 respectively. The 25th, 50th, and 75th percentiles of the cost scaling factor for AI-focused HPC data centres (AI) are 2.03, 11.37, and 40.48 respectively.

A is the price reduction coefficient that represents the proportionate price reduction in response to a certain proportion of job delay. R is the hourly price of the whole CPU (for general-purpose HPCs) or a single GPU (for AI-focused HPCs). Although a resource is rent by vCPU, we use the whole CPU data to have values with magnitude comparable to GPUs. G is the power of a single CPU or a single GPU. These parameters are interpreted in Section 6.1.2. The power of a vCPU (G) is calculated by dividing the rated power of the physical CPU by the number of vCPUs. Note that, in the plots of R , G , and R/G , when a set of computing rental options are essentially renting different portions of the same type of machine, this chapter only keeps one option to avoid over-counting the same machine for its power and price parameter.

To evaluate the financial profitability of providing power system services, the prices of power system services in the UK, Australia, and Germany are collected in this chapter. In the winters of 2022/23 and 2023/24, the UK's Demand Flexibility Service (DFS) offered a guaranteed acceptance price at 3 GBP/kWh (around 3.8 USD/kWh) for flexibility during stressful system periods [419] (note that the high price may reflect its status as a trial programme intended to incentivise participation, and could be lower once formally launched). 34 test events and 4 live events were performed. For Australia, the price data from the National Energy Market (NEM) for 2022–2024 [418] are collected. For Germany, the 2022–2024 prices of automatic frequency restoration reserve (aFRR) are collected, where aFRR is a secondary reserve power system service common among Europe [417]. Fig. 6.2g shows the percentiles of these prices. Comparing the price and cost results, it can be seen that the AI-focused HPC data centre can achieve profitability under the following conditions: 1) when the cost scaling factor is at the 25th percentile and prices are at or above the 99.9th percentiles for aFRR, NEM, and UK DFS; or 2) when the cost scaling factor is at the 50th percentile and prices are at the 99.99th percentile for aFRR and at or above the 99.95th percentile for NEM. Although rare, these price events underscore the high value of certain power system services, which could be profitable for data centres to provide.

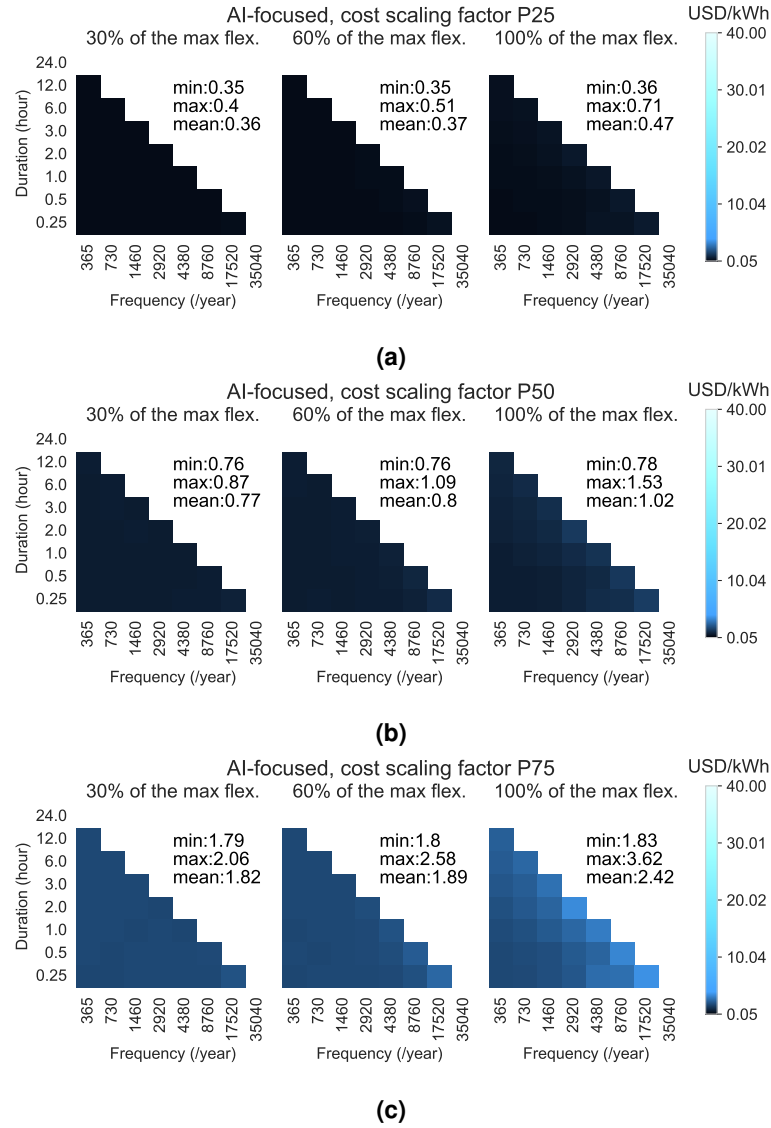


Figure 6.4: Average flexibility cost under the cost scaling factor derived by Lambda GPU Cloud data.

These cost results are for the AI-focused HPC data centre (Saturn) when providing different percentages of the maximum amounts of flexibility evaluated in Fig. 6.1c. The maximum delay limit is set to 20%. “min”, “max”, and “mean” refer to the minimum, maximum, and mean values across all blocks in each heatmap. (a), (b), and (c) are for 25-th, 50-th, and 75-th percentiles of the cost scaling factor respectively. The Lambda GPU Cloud data is sourced from [420].

When gathering data from various cloud providers, it was found that Lambda GPU Cloud [420], which specialises in GPU rentals, offers GPU options approximately 70% cheaper than those from Google Cloud, AWS, and Oracle. The flexibility cost associated with the AI-focused HPC data centre is consequently reduced when using a cost scaling factor based on Lambda GPU Cloud data, as shown in Fig. 6.4. By comparing Fig. 6.4 and Fig. 6.2g, it can be seen that the profitability of the AI-focused HPC data centre can be achieved even under the 75th percentile of the cost scaling factor. This finding indicates that the profitability varies between data centres with different perceived value in computing. Data centres with lower perceived value in computing might achieve higher profits in providing power system services.

6.2.3 Generalisability of the finding

To generalise the previous findings, this section analyses the 7 AI-focused HPC data centres and 7 general-purpose HPC data centres in Table 6.1 for two typical power system services: One is primary response characterised by short duration and high frequency (0.25 hours and 2920 times a year), while the other is congestion management characterised by long duration and low frequency (2 hours and 365 times a year). It should be noted that these two power system services may have different duration and frequency requirements within a certain range. The specific ranges for these requirements are detailed in [344].

Figs. 6.5a and 6.5b display the utilisation time series of the 14 HPC data centres. Fig. 6.5c shows the maximum amount of flexibility versus the average flexibility cost when providing primary response, while Fig. 6.5d shows that for providing congestion management. When providing primary response, it can be seen that AI-focused HPC data centres have lower flexibility cost than all general-purpose HPC data centres. However, it should also be noted that there are two general-purpose HPC data centre (ORNL and CIEMAT) that have maximum flexibility approximately equal to or greater than some AI-focused HPC data centres. When providing congestion management, AI-focused HPC data centres are more advantageous: All AI-focused HPC data centres have at least 50% lower flexibility cost than all general-purpose HPC data centres, and only one AI-focused HPC data centre (PAI) shows approximately 7% less flexibility than a specific general-purpose HPC data centre (MetaCen).

It should be noted that Figs. 6.5c and 6.5d are based on the 50th percentile of the cost scaling factor. Appendix C.4 provides plots under the 25th and 75th percentiles, and the conclusion of this section remains unchanged.

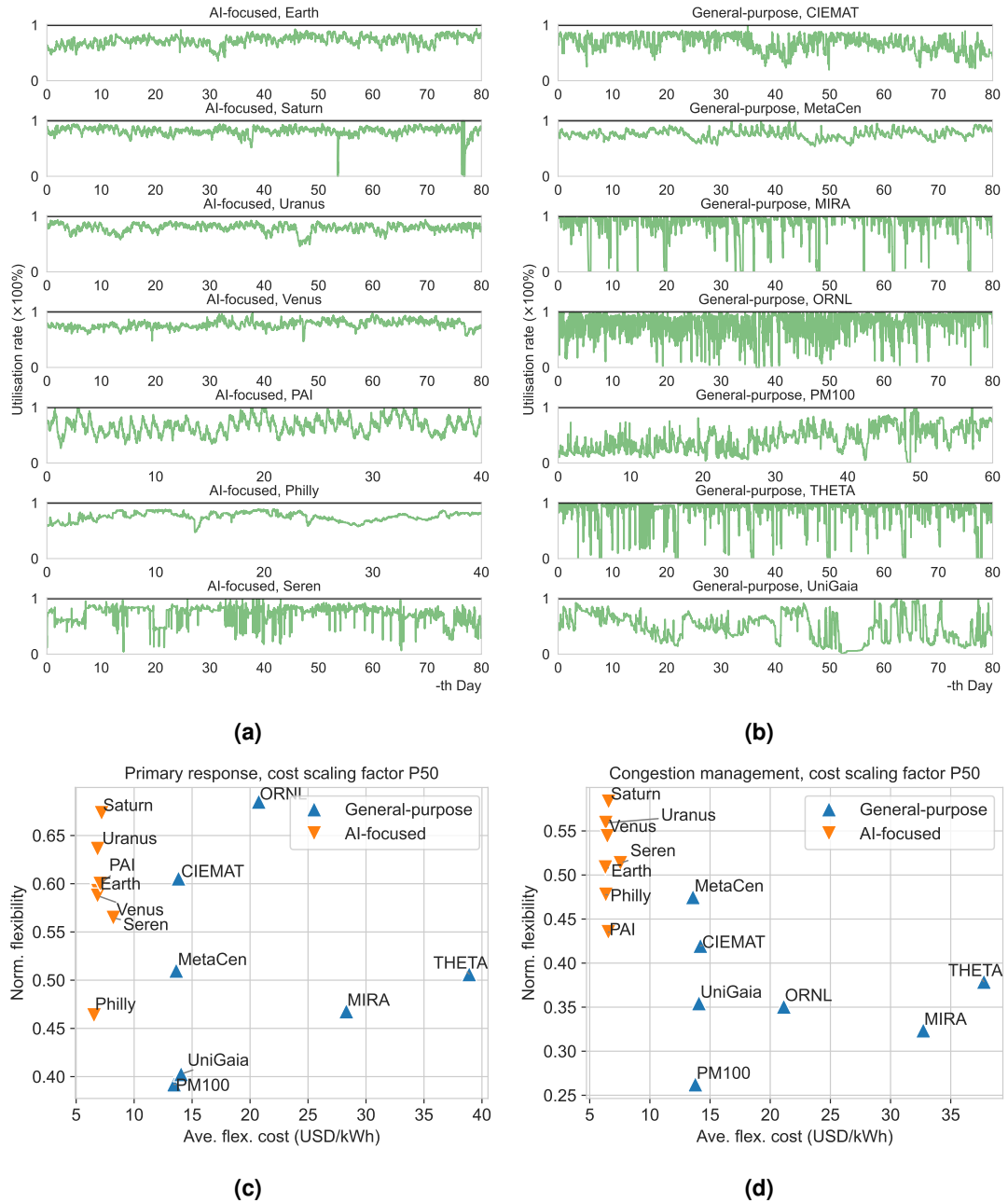


Figure 6.5: Analysis of all the 14 HPC data centres.

(a) Baseline utilisation time series (green lines) of the 7 AI-focused HPC data centres. (b) Baseline utilisation time series (green lines) of the 7 general-purpose HPC data centres. The black lines refer to the utilisation upper bound.

(c) and (d): The normalised maximum amount of flexibility (Norm. flexibility) versus the average flexibility cost (ave. flex. cost) when providing 100% of the maximum flexibility for two power system services. This thesis uses the 50-th percentile (P50) of the cost scaling factor (estimated using data from Google Cloud, AWS, and Oracle). (c) Results for providing primary response (duration of 0.25 hours and frequency of 2920 times/year); (d) Results for providing congestion management (duration of 2 hours and frequency of 365 times/year). Dots closer to the top-left corner indicate better flexibility providers in terms of greater flexibility and lower cost.

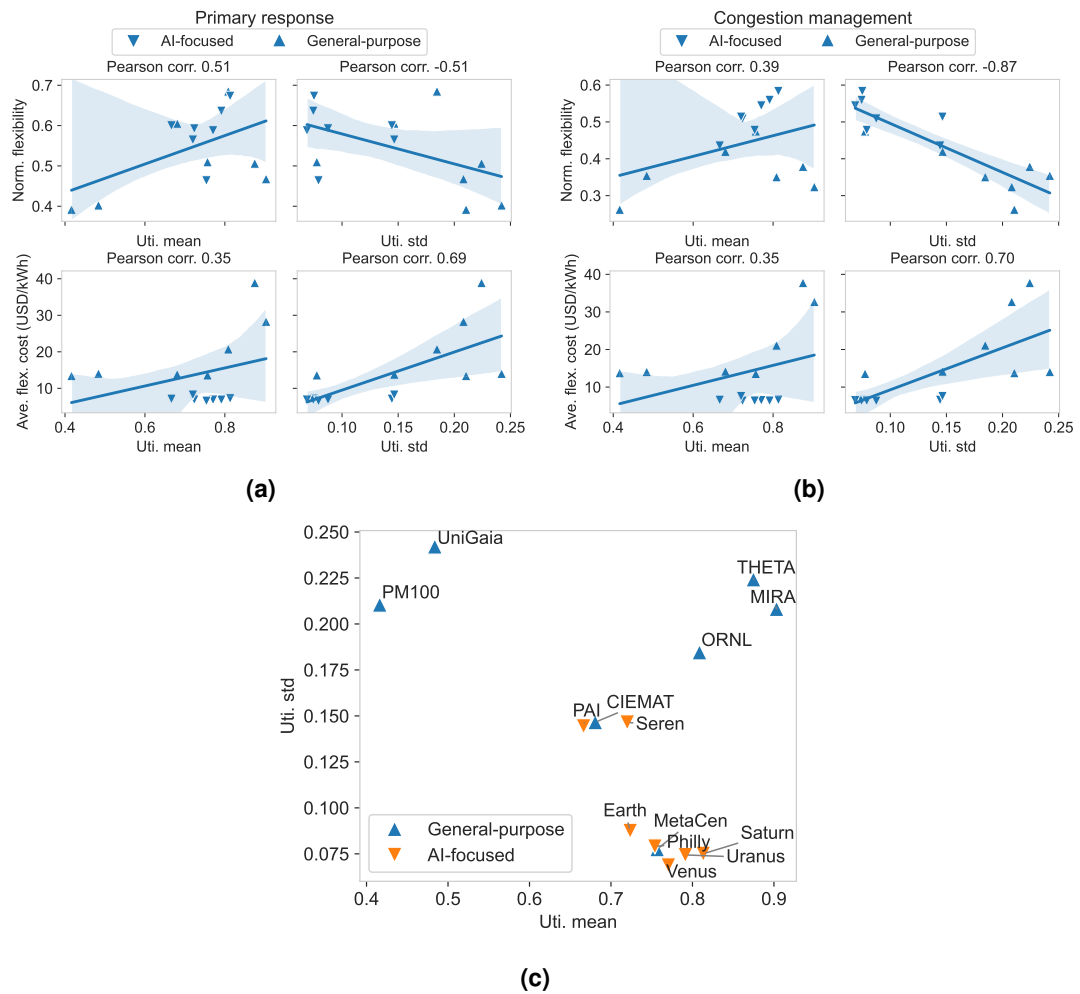


Figure 6.6: Correlation analysis of the flexibility results and data centre utilisation patterns.

The cost scaling factor is set to the 50th percentile (estimated using data from Google Cloud, AWS, and Oracle). (a) Correlation between the normalised maximum amount of flexibility (Norm. flexibility) for primary response, the average flexibility cost (Ave. flex. cost) for primary response, the mean utilisation rate (Uti. mean), and the standard deviation of utilisation (Uti. std). Regression lines with confidence intervals (shaded areas) are plotted for highlighting the trends. (b) Correlation results where the flexibility and cost are for congestion management. (c) The mean utilisation rate and the standard deviation of utilisation for all the 14 data centres.

6.2.4 Impact of job patterns on the maximum flexibility and flexibility cost

In addition to the different cost scaling factors illustrated in Fig. 6.3, the different utilisation patterns also contribute to the greater flexibility and lower cost of AI-focused data centres compared to general-purpose data centres. Fig. 6.6a illustrates the correlation between the normalised maximum amount of flexibility for the primary response service, the average flexibility cost for the primary response service, the mean baseline utilisation rate and the standard deviation of baseline utilisation; Fig. 6.6b shows that for the congestion management service. The first column of Figs. 6.6a and 6.6b shows a positive correlation between the mean baseline utilisation and the normalised maximum amount of flexibility, and between the mean baseline utilisation and the average flexibility cost. This is intuitive as a higher mean utilisation rate implies a higher baseline power, leading to greater flexibility. At the same time, a highly utilised data centre faces more difficulty in restoring jobs disrupted by flexibility provision, resulting in higher cost.

The second column of Figs. 6.6a and 6.6b shows a negative correlation between the standard deviation of utilisation and the normalised maximum amount of flexibility, which is strong (-0.87) for the congestion management service. This is because data centres with variable utilisation time series (indicating variable power demand) are less capable of sustaining flexibility for a long time. In addition, a strong positive correlation is observed between the standard deviation of utilisation and the average flexibility cost. This is because variant utilisation time series can have more frequent utilisation spikes, which impede timely job restoration and thus increase flexibility cost.

Finally, Fig. 6.6c shows the mean utilisation and the standard deviation of utilisation for all data centres. AI-focused HPC data centres tend to have relatively high utilisation rates but low variance, enabling them to provide greater flexibility at lower cost especially for long-duration services such as congestion management, as has been illustrated in Fig. 6.5d. Note that this also highlights an opportunity for general-purpose HPC data centres to improve their flexibility by adjusting their baseline utilisation patterns.

Figs. 6.6a and 6.6b use the 50th percentile of the cost scaling factor. Appendix C.4 provides figures under the 25th and 75th percentiles. The same conclusion can still be derived.

6.2.5 Dynamic quota for greater flexibility and lower cost

This section presents the result under the dynamic quota opportunity, as is modelled in Section 6.1.3. Note that this section adopts the same baseline operation profiles as in the previous sections, which may slightly overestimate the additional flexibility under dynamic quota. However, the dynamic quota remains available as long as the data centre is not operating at full capacity at all times. Figs. 6.7a and 6.7b show the maximum amounts of flexibility for the two data centres in Figs. 6.1c and 6.1d, where 100% extra computing resources can

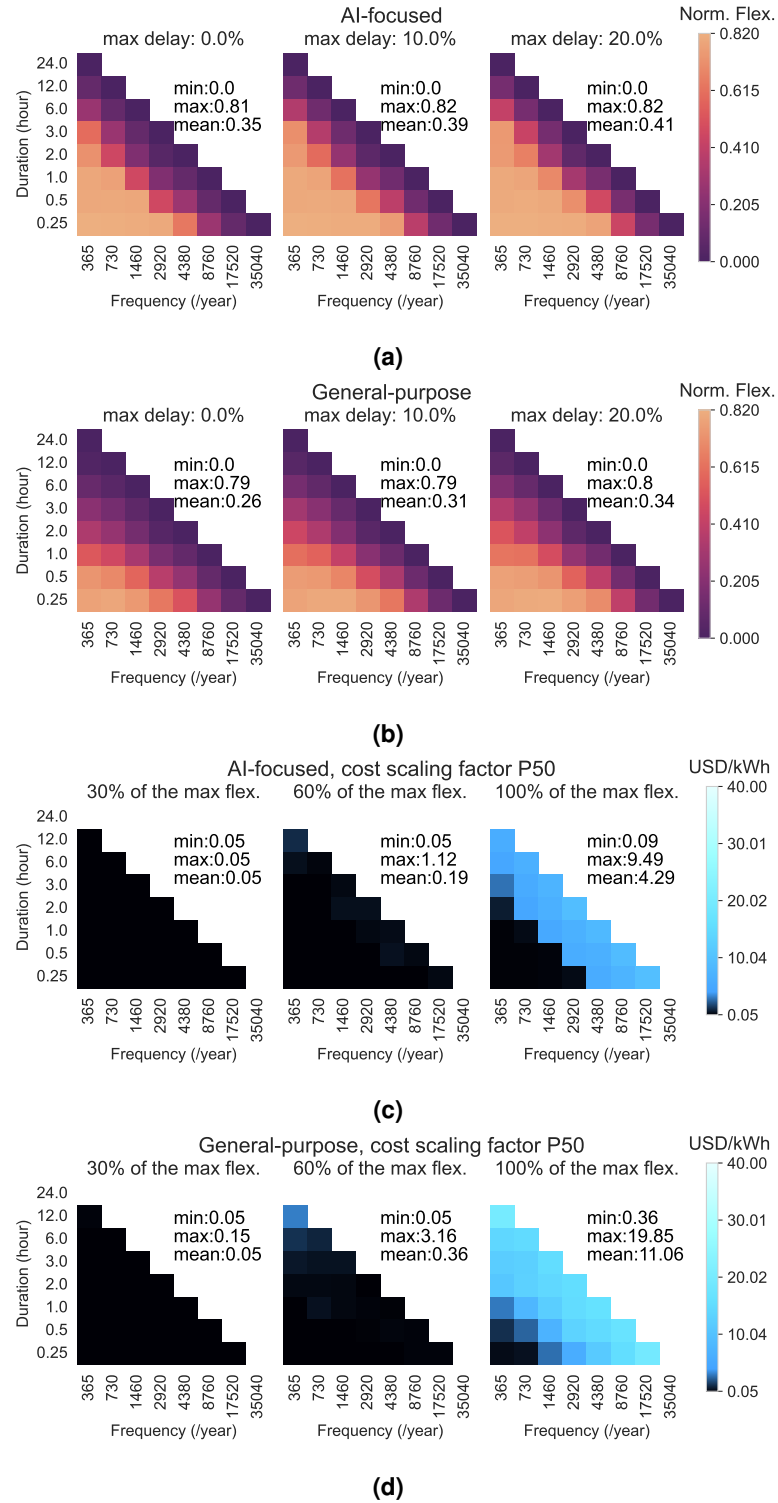


Figure 6.7: The maximum amount of flexibility and the flexibility cost with the dynamic quota opportunity.

(a) The maximum amounts of flexibility for the AI-focused HPC data centre (Saturn). (b) The maximum amounts of flexibility for the general-purpose HPC data centre (ORNL). (c) and (d): The average cost of providing different percentages of the maximum amounts of flexibility evaluated in (a) and (b). (c) The average flexibility cost for the AI-focused HPC data centre (Saturn). (d) The average flexibility cost for the general-purpose HPC data centre (ORNL). In this thesis, the 50-th percentile of the cost scaling factor is used, which is estimated using data from Google Cloud, AWS, and Oracle. “min”, “max”, and “mean” refer to the minimum, maximum, and mean values across all blocks in each subplot. Using 100% extra computing resources is assumed to have 50% computing speed-up.

only increase 50% of computing speed. It can be seen that the maximum flexibility increases noticeably compared to Figs. 6.1c and 6.1d, and there is now significant flexibility even when no computing delay is allowed (this is in contrast to the results without dynamic quota in Figs. 6.1c and 6.1d where there is no flexibility when no delay is allowed).

Figs. 6.7c and 6.7d show the flexibility cost with the dynamic quota opportunity. Using extra resources for computational speed-up imposes additional flexibility cost due to higher energy usage, which are calculated based on a wholesale energy market price of 0.05 USD/kWh. Figs. 6.7c and 6.7d show that the average flexibility cost can be as low as 0.05 USD/kWh, which is close to grid energy storage [344]. This is because data centres with dynamic quota can provide power system services without incurring job delays.

6.2.6 Discussion on Cooling Systems

The flexibility results in this chapter focus primarily on the data centre computing system. While the computing system is the main power consumer within a data centre, the cooling system also accounts for a non-trivial portion of total energy use. Although Google reports an average Power Usage Effectiveness (PUE)—defined as the ratio of total facility energy consumption (including cooling) to IT energy consumption—of 1.09 across its global data centre fleet [421], the industry average remains considerably higher, with cooling systems accounting for up to 30% of total power consumption [421, 422]. Due to thermal inertia, data centre cooling systems can also offer additional flexibility, which, when combined with the flexibility of the computing system, can enable the provision of a larger volume of flexibility services.

6.3 Conclusion

This chapter addresses the last research question on the flexibility and cost of AI-focused data centres. It proposes methods and then provides a detailed analysis of the capability and cost of grid flexibility provision by GPU-heavy, AI-focused HPC data centres, with a comparison to CPU-heavy, general-purpose HPC data centres traditionally used for scientific computing. Using real-world computing traces from 7 AI-focused HPC data centres and 7 general-purpose HPC data centres, along with computing prices from three cloud platforms, the analysis demonstrates that AI-focused HPC data centres can deliver greater flexibility at 50% lower cost than general-purpose HPC data centres across a range of power system services. Furthermore, by comparing the flexibility costs to market prices, the financial profitability of flexibility provision for AI-focused HPC data centres is illustrated.

The findings in this chapter have implications for several stakeholders. For power system operators who want to manage increasing electricity demand while deferring expensive grid infrastructure upgrades, they can design more targeted collaboration strategies with data centres based on their differing capabilities for providing specific power system services. For example, AI-focused HPC data centres can be prioritised when securing power system services that require long duration. The findings here also have implications for data centre operators, who may believe that they should not disrupt “high-value” computing jobs to provide “low-value” power system services. This chapter shows that providing power system services could bring extra net profit to data centres. It is worth noting that the value of flexibility can persist even when data centres are constructed in areas with sufficient connection capacity, because data centre flexibility can support the grid by delivering essential system-wide services like reserves and frequency regulation (e.g., aFRR). Finally, for regulators and policymakers, the data centre flexibility and the cost model in this chapter can help assess the benefits of integrating data centres into power system flexibility markets.

Conclusion

In the context of global decarbonisation, DERs are being rapidly integrated into power systems. While this trend places increasing strain on grid capacity, DERs also offer valuable flexibility that can support stable power system operation. However, fully unlocking this valuable flexibility entails several major challenges, such as those in scheduling DERs under uncertainty, and those in understanding their impacts on fuel poverty and power systems. This thesis addresses these challenges by proposing a suite of novel methods that are validated using a broad range of datasets and case studies.

7.1 Summary of the thesis

The core content of this thesis revolves around addressing key research questions associated with unlocking DER flexibility. These are:

- *Q1: How can flexible DERs be scheduled in a reliable and computationally scalable manner considering the DER uncertainty?*
- *Q2: What are the fuel poverty and distribution network impacts of replacing domestic off-gas heating systems with HPs?*
- *Q3: Do AI-focused data centres have a different potential for power system flexibility compared to general-purpose data centres, and what is their cost of flexibility provision?*

The key findings and contributions in addressing these questions are summarised below:

Uncertainty in scheduling distributed energy resources

Most DERs are inherently characterised by significant uncertainties, stemming from their dependence on external and often unpredictable factors. When these uncertainties are incorporated into the operational constraints of power systems, it is often desirable to use a JCC to ensure the simultaneous satisfaction of multiple constraints with a high probability. However, integrating JCCs into power system operations is computationally challenging, which has limited their adoption in practical industry-scale problems. In response to this challenge and namely the first research question:

- Chapter 3 of this thesis addressed the computational challenges associated with RHS-WDRJCC, a regime that provides better out-of-sample performance than classical JCC and features a special RHS structure. A Strengthened and Faster Linear Approximation (SFLA) was proposed, with theoretical proof demonstrating its ability to be less conservative compared to an existing inner approximation that is equivalent to W-CVaR under certain hyper-parameters. The effectiveness of the proposed SFLA was demonstrated through two practical power system applications. In power system unit commitment (UC) problems, the proposed SFLA achieved up to a 10× speedup and an average 3.8× speedup in finding comparable high-quality feasible solutions compared to the existing strengthened and exact MIP reformulation. In bilevel energy storage strategic bidding problems where the exact formulation is not applicable due to its non-convexity, the SFLA demonstrated a remarkable 90× speed-up compared to W-CVaR.
- DER aggregation is a promising hierarchical scheduling regime that alleviates the computational challenges associated with directly scheduling large numbers of individual DERs at the whole-system level. However, this scheduling approach does not eliminate the inherent uncertainty of DERs, thereby necessitating the integration of uncertainty management techniques, such as JCCs, into DER aggregation. Chapter 4 introduced a novel surrogate-polytope approximation method to efficiently incorporate WDRJCC and JCC with a two-stage structure into DER aggregation. The surrogate-polytope approximation results in a special single-level RHS structure that only applies to the low-dimensional aggregated power. Specifically, based on the proposed surrogate-polytope method, 1) when WDRJCC is incorporated into DER aggregation, the aggregation outcome—the Aggregated Feasible Power Region (AFPR)—admits a tractable analytical MILP reformulation; 2) when classical JCC is integrated, while no analytical expression for the AFPR can be derived, the jointly reliable dispatch of the aggregated power can be computed using an iterative process.

Furthermore, due to the proposed surrogate-polytope method, the integration of both WDRJCC and JCC introduces minimum extra computational complexity to the calculation of AFPR in a deterministic setting. This extra computational complexity does not increase with the number of DERs nor with the size of network. Chapter 4 also proved that the proposed surrogate-polytope method leads to guaranteed inner-approximation to the jointly reliable AFPR defined by the original large-scale two-stage JCC that includes all DER variables, DER constraints, and network constraints. Finally, extensive case studies demonstrated that the proposed method produces a less conservative yet reliable AFPR, making it a practical and efficient approach for addressing uncertainty in DER aggregation.

Unrecognised impact of distributed energy resources

Beyond the technical aspect of managing the computational complexity of DER scheduling under uncertainty, as discussed in Section 2.3, substantial existing work has been done on evaluating the technical, economic and social impacts of many DERs, such as EVs, distributed RES, and distributed ES. These evaluation also represents a crucial dimension that supports the design of effective and efficient roadmaps to realise DER flexibility. However, two important types of DERs, namely HPs and data centres, remain under-investigated. In particular, the impacts of HPs on fuel poverty and distribution networks, as well as the potential for power system flexibility from the emerging AI-focused data centres, are not well understood. Given the pivotal role that HPs and data centres are expected to play in future power systems, this knowledge gap poses a significant challenge in unlocking the full value of DER flexibility. In response to this challenge and namely the second and third research questions, this thesis proposed novel methods and conducted extensive evaluations, leading to the following key findings:

- HPs are central to heating decarbonisation, and the 300% running efficiency of HPs makes them particularly financially attractive for off-gas homes that currently rely on costly heating solutions such as ERH and heating oil. However, it remains unclear whether the introduction of HPs can effectively reduce heating costs for these off-gas homes, alleviate fuel poverty, and be accommodated by the existing distribution network. Chapter 5 addressed this gap through case studies conducted in regions of Scotland and England. The analysis demonstrated that replacing off-gas heating with HPs can reduce fuel poverty during both normal and energy-crisis periods, assuming the availability of existing UK government support for HP upfront costs. However, certain regions are shown to face significant distribution network upgrade cost. This underscores the substantial social benefits of HP rollout while also highlighting the urgent need for HP flexibility to mitigate network stress.
- Data centres are a major source of rapid load growth, driven by the expanding AI sector. However, there is limited understanding of: 1) whether AI-focused data centres can impact power systems by offering a new and significant source of flexibility, and 2) whether data centre operators can be incentivised to offer this flexibility, given that providing power system flexibility may compromise the computing performance of data centres.

Chapter 6 addressed these gaps by analysing and comparing the grid flexibility of 7 AI-focused, GPU-intensive data centres with 7 traditional, CPU-intensive general-purpose data centres. Furthermore, Chapter 6 proposed a method for modelling the cost incurred by data centres in providing these grid flexibility services, incorporating the value of computing by utilising data from three major cloud computing platforms: Google Cloud, Amazon AWS, and Oracle. The analysis revealed that AI-focused HPC data centres can provide greater power flexibility at 50% lower cost than general-

purpose HPC data centres for a range of power system services. This finding has significant implications for both data centre operators and power system operators, motivating greater collaboration to develop flexibility schemes tailored to different types of data centres. By comparing flexibility cost with real-world power system service prices, Chapter 6 also demonstrated the financial profitability of flexibility provision for AI-focused HPC data centres. This challenges the view that computing jobs are always more valuable than power system services, providing financial motivation for data centre operators to participate in grid flexibility activities.

In summary, this thesis has provided novel methodologies and actionable insights relevant for the integration of DER flexibility into power systems. These contributions are expected to support the development of more resilient, reliable, and sustainable power systems in the context of the ongoing penetration of DERs.

7.2 Future work

The work in this thesis also motivates several future research directions towards the full utilisation of DER flexibility.

7.2.1 Scheduling distributed energy resources with a multi-stage uncertainty structure

A promising direction for future work lies in extending the proposed SFLA (Chapter 3) to multi-stage stochastic programming with JCC/WDRJCC for DER scheduling. As discussed in Sections 4.7.2 and 4.7.3, real-world DER scheduling can have continuous adjustments to power dispatch as uncertainty is gradually observed, which helps achieve the lowest possible operation cost. The multi-stage stochastic programming framework with JCC/WDRJCC would ensure that, for the here-and-now decision, there is a high probability that feasible future DER operational strategies can be found to satisfy all constraints as uncertainty unfolds. This framework closely mirrors the real-world dispatch problem and, as such, represents the ideal performance of DER scheduling. However, such a framework has limited practical application in DER scheduling due to the excessive computational complexity. Valid inequalities used for the proposed SFLA represent a promising solution to this computational issue. Our SFLA in Chapter 3 has demonstrated the power of valid inequalities in achieving around 100x computational speedup without compromising solution quality. It is therefore worth investigating the possibility of extending the benefits to the multi-stage JCC/WDRJCC for DER flexibility scheduling problems.

A related future avenue is the study of JCC/WDRJCC with a left-hand-side (LHS) uncertainty structure (see Section 2.2.1). Although the RHS structure studied in Chapter 3 can handle several practical DER scheduling problems, such as managing the thermal constraints or securing the power system reserve, the LHS structure is more general and can facilitate more real-world applications. A typical example is the multi-stage DER scheduling problem as discussed above. In a multi-stage problem with a linear decision rule, the LHS-JCC will naturally arise [145]. Valid inequalities used to derive the proposed SFLA can be investigated as well. Furthermore, the proposed SFLA may be applied in a natural way, by noticing the fact that the LHS uncertainty reduces to the RHS uncertainty for any specific choice of decision variables that affect the LHS coefficients.

Similar to what has been investigated in this thesis, this future direction also involves integrating the multi-stage setting into the DER aggregation process. By integrating DER aggregation, this extension would also alleviate the computational complexity for the whole-system DER scheduling with the multi-stage uncertainty structure when DERs are integrated at scale. As has been extensively discussed in Section 4.7.2, this remains an open question in the existing literature on DER aggregation and presents a promising opportunity for advancing the field. The proposed SFLA and also the use of valid inequalities (as discussed above) can also be exploited to speed up the calculation of AFPR under such multi-stage uncertainty structure.

Beyond the theoretical advancements in solving more complex uncertainty-aware power system problems, a promising direction for future work lies in improving the interpretability and trustworthiness of the chance-constrained framework. Although mathematically appealing, power system operators may be reluctant to rely on a safety guarantee that holds only with a certain probability. They may find robust optimisation (RO) more attractive, as RO offers guarantees against the worst-case scenario—even if that “worst-case” is defined within a prescribed uncertainty set and is not the true worst case. A potential solution is to establish a relationship or equivalence between the probability level in JCC and the size of the uncertainty set in RO. This could help operators better understand the nature of the safety guarantees provided by JCC, thereby improving their trust in the framework.

Issues of interpretability and trustworthiness also arise in market-clearing problems using JCC, where it may be unclear why a certain low-cost generator is not dispatched. Addressing this issue may also require linking the JCC framework to the existing interpretative mechanisms used in current market-clearing processes, in order to clearly explain to affected generators why they are not selected.

7.2.2 Understanding the potential and value of heat pump flexibility

Chapter 5 of this thesis demonstrates both the fuel poverty benefits and the distribution network impacts (significant for certain UK regions) brought by HPs. Specifically, the network impacts underscore the need for HP flexibility, and thus a natural extension is to evaluate the extent to which HP flexibility can be activated and the resulting reduction in network costs. This analysis provides valuable guidelines for policy design related to HP flexibility and supports successful project applications for real-world HP flexibility deployment. This analysis may involve various factors from both technical and social perspectives. From a technical perspective, fully utilising HP flexibility is challenging due to the large problem scale and inherent uncertainty, which are linked to the computational challenges previously discussed for the JCC integration. A robust evaluation of HP flexibility should account for these computational challenges alongside current bottlenecks in software and hardware, ensuring that the results reflect the practically achievable value of HP flexibility. From a social perspective, since HPs are closely tied to residential users, the activation of flexibility also depends on users' willingness to engage. This aspect needs to be included to avoid an overestimate of the value of HP flexibility. Accounting for this factor may require insights from the social sciences and real-world trial programmes for HP flexibility. The above extended analysis can be integrated into the proposed HP evaluation framework for fuel poverty and network upgrade costs, alongside the supporting datasets on GB households and distribution networks collected in this thesis.

7.2.3 Realising power system flexibility of AI-focused data centres

This thesis highlighted the greater potential of AI-focused data centres in providing power grid flexibility. Based on this finding, a future direction is to realise this flexibility with more practical considerations. Similar to HPs, realising data centre flexibility also needs to consider the computational bottlenecks in the real-time flexibility dispatch. The real-time implementation of flexible data centres is challenging due to the large number of computing jobs and the necessary uncertainty management required to ensure reliable delivery of these jobs, as governed by SLA (see Section 2.3.5). Here the computational bottlenecks are further exacerbated by the fact that the data centre flexibility utilisation must avoid excessive use of computing resources that could otherwise be allocated to high-value computing tasks. The work presented in this thesis offers several avenues of support for addressing these challenges. On the one hand, the proposed SFLA and the use of valid inequalities may be explored to mitigate computational bottlenecks in real-time data centre flexibility delivery. On the other hand, the extensive flexibility analysis conducted in this thesis identifies power system services that offer high potential and low cost, thereby providing a practical guideline for prioritising the realisation of data centre flexibility for specific grid services.

Appendix for the Strengthened and Faster Linear Approximation

This appendix provides the detailed mathematical proofs, case study settings, and supplementary analysis for Chapter 3.

A.1 Strengthened and exact MIP reformulation

The strengthened exact MIP reformulation (ExactS), presented as Formulation (20) in [108], is reproduced below for convenience:

$$\mathbf{z} \in \{0, 1\}^N, s \geq 0, \mathbf{r} \geq \mathbf{0}, \quad (\text{A.1a})$$

$$\epsilon N s - \sum_{i \in [N]} r_i \geq \theta N, \quad (\text{A.1b})$$

$$\sum_{i \in [N]} z_i \leq \lfloor \epsilon N \rfloor, \quad (\text{A.1c})$$

$$M(1 - z_i) \geq s - r_i, \quad \forall i \in [N], \quad (\text{A.1d})$$

$$\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} + \frac{-\mathbf{b}_p^\top \boldsymbol{\xi}_i + q_p}{\|\mathbf{b}_p\|_*} z_i \geq s - r_i, \quad \forall i \in [N], p \in [P], \quad (\text{A.1e})$$

$$\frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq s, \quad \forall p \in [P]. \quad (\text{A.1f})$$

where all the notations have been defined in the main content. The $\mathbf{x}$ -feasible region of the ExactS can be expressed as:

$$\mathcal{X}_{\text{ExactS}} = \{\mathbf{x} \in \mathcal{X} \mid \exists \mathbf{z}, s, \mathbf{r} : (\text{A.1a})\text{--}(\text{A.1f})\}. \quad (\text{A.2})$$

A.2 Proofs

Proof of Proposition 1

Proof. Note that the first two constraints in sets (3.9) and (3.20) are identical. To show the statement, it suffices to show that replacing the constraint (3.7c) in set (3.9) with constraints (3.19c)–(3.19d) in set (3.20) has no impact on the $\mathbf{x}$ -feasible region.

(3.20) $\implies$ (3.9): Based on the definition of $[N]_p$, it follows that:

$$\mathbf{b}_p^\top \boldsymbol{\xi}_i \geq q_p, \quad \forall i \in [N] \setminus [N]_p, \quad p \in [P].$$

Then, based on (3.19d), for any $i \in [N] \setminus [N]_p$ with $p \in [P]$, it follows that:

$$\left(\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right)^+ \geq \frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq \frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq s.$$

As $r_i \geq 0$, the following can be deduced:

$$\left(\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right)^+ \geq s \geq s - r_i,$$

which, combined with (3.19c), implies the satisfaction of (3.7c).

(3.20) $\Leftarrow$ (3.9): Based on Lemma 1, for any $\mathbf{x} \in X_{\text{Exact}}$ in (3.9), there always exists $(\mathbf{r}^*, s^*)$ such that s^* is equal to the $(k+1)$ -th smallest value amongst the set $\{\text{dist}(\boldsymbol{\xi}_i, \mathcal{S}(\mathbf{x}))\}_{i \in [N]}$ and constraints (3.7a)–(3.7c) are satisfied. Without loss of generality, the thesis assumes the following ordering: $\text{dist}(\boldsymbol{\xi}_1, \mathcal{S}(\mathbf{x})) \leq \dots \leq \text{dist}(\boldsymbol{\xi}_N, \mathcal{S}(\mathbf{x}))$. Then denote the $(k+1)$ -th smallest distance value as $\text{dist}^*(\mathbf{x}) := \text{dist}(\boldsymbol{\xi}_{k+1}, \mathcal{S}(\mathbf{x})) = s^*$. Combined with the analytical expression of the distance function in (3.6), the pair $(\mathbf{x}, \mathbf{r}^*, s^*)$ satisfies the constraints (3.19a)–(3.19c). It remains to show the satisfaction of (3.19d) given the pair $(\mathbf{x}, \mathbf{r}^*, s^*)$. Now consider two cases:

- If $\text{dist}^*(\mathbf{x}) = 0$, then $s^* = 0$. As (3.7b) must be satisfied for the given $(\mathbf{r}^*, s^*)$, this leads to a contradiction, because $s^* = 0$ implies the non-positivity of the left-hand side of (3.7b), but the right-hand side of (3.7b) is strictly positive due to the assumption $\theta > 0$. Therefore, it can be derived that $\text{dist}^*(\mathbf{x}) > 0$.
- If $\text{dist}^*(\mathbf{x}) > 0$, then for any $p \in [P]$, when $i \geq k+1$, it follows that:

$$0 < s^* = \text{dist}^*(\mathbf{x}) \leq \text{dist}(\boldsymbol{\xi}_i, \mathcal{S}(\mathbf{x}))$$

Given $0 < \text{dist}(\boldsymbol{\xi}_i, \mathcal{S}(\mathbf{x}))$, based on (3.6), it follows that:

$$\text{dist}(\boldsymbol{\xi}_i, \mathcal{S}(\mathbf{x})) = \min_{p \in [P]} \frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*}$$

which leads to:

$$s^* \leq \min_{p \in [P]} \frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \leq \frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*}.$$

In other words, for any $p \in [P]$, there will be at least $N - k$ elements in the set $\left\{ \frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right\}_{i \in [N]}$ not smaller than s^* . Since q_p is defined as the $(k + 1)$ -th smallest value of $\{\mathbf{b}_p^\top \boldsymbol{\xi}_i\}_{i \in [N]}$, $\frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*}$ is the $(k + 1)$ -th smallest element in the set $\left\{ \frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right\}_{i \in [N]}$. Hence, it follows

$$s^* \leq \frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*},$$

which leads to the satisfaction of (3.19d). □

Proof of Theorem 1

Proof. By comparing $\mathcal{X}_{\text{SFLA}}(\kappa)$ in (3.22) and $\mathcal{X}_{\text{Exact}}$ through its equivalent and strengthened reformulation in (3.20), the proof is apparent due to the following inequality given $\kappa_i \in [0, 1]$ for all $i \in [N]$:

$$\left(\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right)^+ \geq \kappa_i \left(\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right).$$

□

Proof of Lemma 2

Proof. (3.23) $\implies$ (3.16): Denote $-\widehat{\text{dist}}(\boldsymbol{\xi}_i, \mathcal{S}(\mathbf{x}))$ by $v(\boldsymbol{\xi}_i)$, and assume that s' is selected such that the term $s' + \frac{1}{\epsilon N} \sum_{i \in [N]} \max\{0, v(\boldsymbol{\xi}_i) - s'\}$ is minimised. Then, take $r_i^* = \max\{0, v(\boldsymbol{\xi}_i) - s'\}$ (implying $r_i^* \geq 0$). The inequality in (3.23) indicates the following:

$$\frac{\theta}{\epsilon} + s' + \frac{1}{\epsilon N} \sum_{i=1}^N r_i^* \leq 0 \Leftrightarrow \epsilon N s' + \sum_{i=1}^N r_i^* \leq -\theta N. \quad (\text{A.3})$$

Taking $s^* = -s'$, now the thesis will prove that the tuple $(\mathbf{x}, s^*, \mathbf{r}^*)$ satisfy all the constraints in (3.16), namely (3.15a)–(3.15c), for any $\mathbf{x}$ belonging to the set (3.23). Since $r_i^* \geq 0$ and $-\theta N \leq 0$, (A.3) implies $s' \leq 0$ (i.e., $s^* \geq 0$) and the following inequality:

$$\epsilon N s - \sum_{i=1}^N r_i \geq \theta N,$$

which means (3.15a) and (3.15b) are satisfied. It remains to prove the satisfaction of (3.15c). Based on the setting of r_i^* , it follows that:

$$s^* - r_i^* = s^* - \max \{0, v(\xi_i) + s^*\} = s^* - \max \{0, s^* - \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\}. \quad (\text{A.4})$$

Notice that constraint (3.15c) is equivalent to:

$$\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})) = \min_{p \in [P]} \left\{ \kappa_i \frac{\mathbf{b}_p^\top \xi + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right\} \geq s - r_i \quad \forall i \in [N]. \quad (\text{A.5})$$

When $\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})) \geq s^*$, (3.15c) must be satisfied for $(\mathbf{x}, s^*, \mathbf{r}^*)$ due to $r_i^* \geq 0$. When $\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})) < s^*$, it can be derived that $s^* - r_i^* = \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))$ based on (A.4). Therefore, (A.5) and equivalently (3.15c) is satisfied.

(3.23) $\Leftarrow$ (3.16): For any $\mathbf{x}$ belonging to the set (3.16), by definition there exist $(\mathbf{r}, s)$ such that constraints in the set (3.16), namely (3.15a)–(3.15c), are satisfied. Among these constraints, (3.15c) (equivalently (A.5)) informs that:

$$r_i \geq s - \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})).$$

Combining with $r_i \geq 0$ in (3.15a), it follows that:

$$r_i \geq \max \{0, s - \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\}.$$

Then the following can be deduced:

$$\epsilon N s - \sum_{i \in [N]} r_i \leq \epsilon N s - \sum_{i \in [N]} \max \{0, s - \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\}.$$

Because of $\epsilon N s - \sum_{i \in [N]} r_i \geq \theta N$ in (3.15b), it follows that:

$$\epsilon N s - \sum_{i \in [N]} \max \{0, s - \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\} \geq \theta N,$$

which means:

$$\frac{\theta}{\epsilon} - s + \frac{1}{\epsilon N} \sum_{i \in [N]} \max \{0, s - \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\} \leq 0.$$

Now take $s' = -s$, it can be deduced that:

$$\frac{\theta}{\epsilon} + s' + \frac{1}{\epsilon N} \sum_{i \in [N]} \max \{0, -s' - \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\} \leq 0$$

which leads to:

$$\frac{\theta}{\epsilon} + \min_{s'} \left\{ s' + \frac{1}{\epsilon N} \sum_{i \in [N]} \max\{0, -s' + v(\xi_i)\} \right\} \leq 0.$$

Therefore, the constraint in (3.23) is satisfied. $\square$

Proof of Lemma 3

Proof. Notice that the CVaR of a random variable $v(\xi)$ with $\xi \sim \mathbb{P}_N$ in (3.24) can be equivalently expressed as the following optimisation form:

$$\mathbb{P}_N\text{-CVaR}_{1-\epsilon}(v(\xi)) = \min_{s' \in \mathbb{R}, \mathbf{r} \geq 0} s' + \frac{1}{\epsilon N} \sum_{i=1}^N r_i \quad (\text{A.6a})$$

$$\text{s.t. } r_i \geq v(\xi_i) - s', \quad i \in [N]. \quad (\text{A.6b})$$

By strong duality of linear programming, the following can be deduced:

$$\mathbb{P}_N\text{-CVaR}_{1-\epsilon}(v(\xi)) = \max_{\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}} \frac{1}{\epsilon N} \sum_{i \in [N]} v(\xi_i) y_i \quad (\text{A.7a})$$

$$\text{s.t. } \frac{1}{N} \sum_{i \in [N]} y_i = \epsilon. \quad (\text{A.7b})$$

Without loss of generality, the thesis assumes the following ordering: $v(\xi_1) \geq \dots \geq v(\xi_N)$. Then a primal-dual optimal pair for CVaR is given by:

$$s'^* = v(\xi_{k+1}), \quad (\text{A.8a})$$

$$r_i^* = \max\{0, v(\xi_i) - s'^*\}, \quad (\text{A.8b})$$

$$y_i^* = \begin{cases} 1, & i = 1, \dots, k \\ \epsilon N - k, & i = k+1 \\ 0, & i > k+1. \end{cases} \quad (\text{A.8c})$$

The validity of the above primal-dual optimal pair holds because of the satisfaction of primal and dual constraints ($\mathbf{y}^* \in [0, 1]$ and $\mathbf{r}^* \geq 0$) and the equality of the objectives of the primal and dual problems:

$$\frac{1}{\epsilon N} \sum_{i \in [N]} v(\xi_i) y_i^* = s'^* + \frac{1}{\epsilon N} \sum_{i \in [N]} r_i^*,$$

which holds because of:

$$\begin{aligned} \frac{1}{\epsilon N} \sum_{i \in [N]} v(\xi_i) y_i^* &= \frac{1}{\epsilon N} \left(\sum_{i \in [k]} v(\xi_i) + v(\xi_{k+1})(\epsilon N - k) \right) \\ &= v(\xi_{k+1}) + \frac{1}{\epsilon N} \left(\sum_{i \in [k]} v(\xi_i) - k v(\xi_{k+1}) \right) \end{aligned}$$

and

$$\begin{aligned} s'^* + \frac{1}{\epsilon N} \sum_{i \in [N]} r_i^* &= v(\xi_{k+1}) + \frac{1}{\epsilon N} \sum_{i \in [N]} \max\{0, v(\xi_i) - v(\xi_{k+1})\} \\ &= v(\xi_{k+1}) + \frac{1}{\epsilon N} \sum_{i \in [k]} (v(\xi_i) - v(\xi_{k+1})) \\ &= v(\xi_{k+1}) + \frac{1}{\epsilon N} \left(\sum_{i \in [k]} v(\xi_i) - k v(\xi_{k+1}) \right), \end{aligned}$$

where the second equality holds because of the ordering of $v(\xi_i)$ such that $v(\xi_i) - v(\xi_{k+1}) \leq 0$ for $i \geq k+1$.

Now take $v(\xi_i) = -\widehat{\text{dist}}(\xi_i, \mathcal{S}(x))$. Based on the CVaR interpretation of $\mathcal{X}_{\text{LA}}(\kappa)$ in (3.23) and the CVaR value expressed as (A.6a), for any $x \in \mathcal{X}_{\text{LA}}(\kappa)$, it can be derived that:

$$\frac{\theta}{\epsilon} + \mathbb{P}_N\text{-CVaR}_{1-\epsilon} \left(-\widehat{\text{dist}}(\xi_i, \mathcal{S}(x)) \right) \leq 0 \quad (\text{A.9a})$$

$$\Leftrightarrow \frac{\theta}{\epsilon} + s'^* + \frac{1}{\epsilon N} \sum_{i=1}^N r_i^* \leq 0 \quad (\text{A.9b})$$

$$\Leftrightarrow \epsilon N s'^* + \sum_{i=1}^N r_i^* \leq -\theta N. \quad (\text{A.9c})$$

Since $r_i^* \geq 0$ and $-\theta N \leq 0$, (A.9c) indicates $s^* := -s'^* \geq 0$ and

$$\epsilon N s^* - \sum_{i \in [N]} r_i^* \geq \theta N, \quad (\text{A.10})$$

which means (3.15a) and (3.15b) are satisfied. It remains to prove the satisfaction of (3.15c). Based on the setting of r_i^* in (A.8b), it follows that:

$$s^* - r_i^* = s^* - \max\{0, v(\xi_i) + s^*\} = s^* - \max\{0, s^* - \widehat{\text{dist}}(\xi_i, \mathcal{S}(x))\}. \quad (\text{A.11})$$

Recall that constraint (3.15c) is equivalent to the following (see (A.5)):

$$\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})) = \min_{p \in [P]} \left\{ \kappa_i \frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right\} \geq s - r_i \quad i \in [N].$$

When $\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})) \geq s^*$, (3.15c) must be satisfied due to $r_i^* \geq 0$. When $\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})) < s^*$, it follows that $s^* - r_i^* = \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))$, so (3.15c) is satisfied as well. Therefore, the thesis has proved that the pair $(\mathbf{x}, s^*, \mathbf{r}^*)$ satisfies constraints in (3.15) for any $\mathbf{x} \in \mathcal{X}_{\text{LA}}(\kappa)$, with $s^* = -s'^* = -v(\xi_{k+1})$ and $r_i^* = \max\{0, v(\xi_i) + s^*\}$. Finally, based on the assumed ordering, $v(\xi_{k+1}) = -\widehat{\text{dist}}(\xi_{k+1}, \mathcal{S}(\mathbf{x}))$ is the $(k+1)$ -th greatest value amongst $\{v(\xi_i)\}_{i \in [N]}$, so $s^* = -v(\xi_{k+1})$ is the $(k+1)$ -th smallest value amongst $\{\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\}_{i \in [N]}$. $\square$

Proof of Theorem 2

Proof. It is equivalent to prove that for any $\mathbf{x} \in \mathcal{X}_{\text{LA}}(\kappa)$, it follows that $\mathbf{x} \in \mathcal{X}_{\text{SFLA}}(\kappa)$, namely that constraints (3.21a)–(3.21d) can be satisfied. Based on Lemma 3, for any $\mathbf{x} \in \mathcal{X}_{\text{LA}}(\kappa)$ in (3.16), there always exists $(\mathbf{r}^*, s^*)$ such that s^* is equal to the $(k+1)$ -th smallest value amongst the set $\{\widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x}))\}_{i \in [N]}$ and constraints (3.15a)–(3.15c) are satisfied, which indicate the satisfaction of (3.21a)–(3.21c).

It remains to show the satisfaction of (3.21d) given the same $(\mathbf{r}^*, s^*)$. Without loss of generality, suppose the following ordering: $\widehat{\text{dist}}(\xi_1, \mathcal{S}(\mathbf{x})) \leq \dots \leq \widehat{\text{dist}}(\xi_N, \mathcal{S}(\mathbf{x}))$. Then denote its $(k+1)$ -th smallest element as $\widehat{\text{dist}}^*(\mathbf{x}) := \widehat{\text{dist}}(\xi_{k+1}, \mathcal{S}(\mathbf{x})) = s^*$. Based on (3.14), it follows that:

$$\widehat{\text{dist}}^*(\mathbf{x}) = \kappa_{k+1} \left(\min_{p \in [P]} \frac{\mathbf{b}_p^\top \xi_{k+1} + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right).$$

Then for any $p \in [P]$, when $i \geq k+1$, it follows that:

$$s^* = \widehat{\text{dist}}^*(\mathbf{x}) \leq \widehat{\text{dist}}(\xi_i, \mathcal{S}(\mathbf{x})) = \kappa_i \left(\min_{p \in [P]} \frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right) \leq \kappa_i \left(\frac{\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \right).$$

That is, for any $p \in [P]$, there will be at least $N - k$ elements not less than s^* for the set $\{\kappa_i (\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}) / \|\mathbf{b}_p\|_*\}_{i \in [N]}$. Since q_p is defined as the $(k+1)$ -th smallest value of $\{\mathbf{b}_p^\top \xi_i\}_{i \in [N]}$, the element $(q_p + d_p - \mathbf{a}_p^\top \mathbf{x}) / \|\mathbf{b}_p\|_*$ is the $(k+1)$ -th smallest element of the set $\{(\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}) / \|\mathbf{b}_p\|_*\}_{i \in [N]}$. Then, it can be deduced that $(q_p + d_p - \mathbf{a}_p^\top \mathbf{x}) / \|\mathbf{b}_p\|_*$ is not less than the $(k+1)$ -th smallest element of the set $\{\kappa_i (\mathbf{b}_p^\top \xi_i + d_p - \mathbf{a}_p^\top \mathbf{x}) / \|\mathbf{b}_p\|_*\}_{i \in [N]}$ due to $\kappa_i \in [0, 1]$. Therefore, it can be deduced that:

$$s^* \leq \frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*}$$

which indicates the satisfaction of (3.21d). $\square$

Proof of Corollary 1

Proof. Proposition 9 in [122] demonstrates that the LA method can achieve an optimal value equivalent to the exact reformulation of RHS-WDRJCC by optimizing κ . Then, the objective function $c(\cdot)$ and the set $\mathcal{X}$ can be adjusted to identify each individual feasible solution $\mathbf{x} \in \mathcal{X}_{\text{Exact}}$, thereby establishing set equivalence. This set equivalence also applies to the proposed SFLA due to Theorem 2. $\square$

Proof of Corollary 3

Proof. The definition of $[N]_p$ leads to:

$$\mathbf{b}_p^\top \boldsymbol{\xi}_i \geq q_p, \quad \forall i \in [N] \setminus [N]_p, \quad p \in [P]. \quad (\text{A.12})$$

Therefore, the constraint (3.21d) in $\mathcal{X}_{\text{SFLA}}(\mathbf{1})$ implies that for any $i \in [N] \setminus [N]_p$ with $p \in [P]$, the following can be derived:

$$\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq \frac{q_p + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq s. \quad (\text{A.13})$$

As $r_i \geq 0$, it follows that:

$$\frac{\mathbf{b}_p^\top \boldsymbol{\xi}_i + d_p - \mathbf{a}_p^\top \mathbf{x}}{\|\mathbf{b}_p\|_*} \geq s \geq s - r_i \quad (\text{A.14})$$

which, combined with (3.21c) in $\mathcal{X}_{\text{SFLA}}(\mathbf{1})$, implies the satisfaction of (3.15c) when $\kappa = \mathbf{1}$. The above process implies $\mathcal{X}_{\text{SFLA}}(\mathbf{1}) \subseteq \mathcal{X}_{\text{LA}}(\mathbf{1})$. Combining with Theorem 2, it follows that $\mathcal{X}_{\text{SFLA}}(\mathbf{1}) = \mathcal{X}_{\text{LA}}(\mathbf{1})$. $\square$

Proof of Corollaries 4 and 5

Proof. The properties in Corollaries 4 and 5 hold for $\mathcal{X}_{\text{LA}}(\kappa)$ (3.16) as was proved from the perspective of the optimal value in [122]. This perspective can be extended to the whole set following the proof of Corollary 1, which in turn holds for the proposed SFLA $\mathcal{X}_{\text{SFLA}}(\kappa)$ due to Corollary 3. In fact, the first condition in Corollary 5 can be relaxed from $\forall \mathbf{x} \in \mathcal{X}_{\text{Exact}}$ to just the optimal solution $\mathbf{x}^*$, as the primary interest is typically in the optimal solution rather than the entire set. $\square$

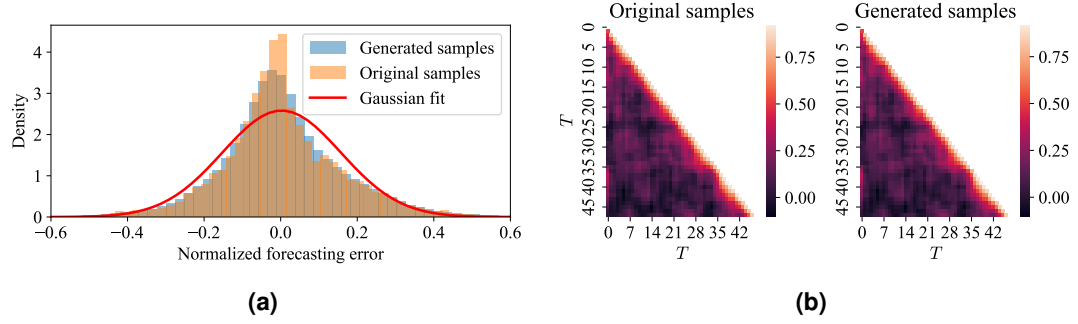


Figure A.1: Comparison of original samples and generated samples for wind forecasting error.

(a) The marginal distribution of the wind forecasting error, which is normalised by the wind generator capacity. Original samples are those 156 samples from Global Energy Forecasting Competition 2012 - Wind Forecasting, with each sample representing the next-48-hour forecasts. Generated samples refer to those generated by the fitted Gaussian KDE plus copula. A Gaussian fit is plotted to highlight the non-Gaussian characteristic of both the original samples and the generated samples. (b) The temporal Pearson correlation coefficients.

A.3 Supplementary information for the unit commitment problems

A.3.1 Case study settings for the unit commitment problem

The UC problem is implemented on the IEEE 24-bus network [423]. The number of gas generators in this network is set to 100. The capacity of individual gas generators varies between 7.6 MW and 17.8 MW, and the total gas generation capacity is 1272 MW that follows the IEEE 24-bus test case data. The generator minimum uptime is set to 2 hours and the minimum downtime is set to 1 hour, representing gas turbines [77]. Other generator technical parameters follow [77]. The case study considers 10 wind farms with a total capacity of 730 MW, which is 60% of the total network load and represents a high level of renewable penetration. Following [126], this thesis considers uncertain wind power generation, which is modelled as a deterministic forecast plus a random forecasting error. Wind data, including forecasts and error samples, are obtained from open-access Bronze-medal forecasts (by Mudit Gaur) for the second wind farm in the Global Energy Forecasting Competition 2012 – Wind Forecasting [424]. The competition task is to forecast the wind power generation for the next 48 hours based on real-world wind farm data. The thesis takes the out-of-sample forecasts from the Bronze-medal forecasting results, and then calculates the difference to the released ground-truth [425] to get 156 samples of wind power forecasting error, with each sample being the 48-hour forecasting error. These samples enable the calculation of the temporal correlation matrix and the marginal distribution of the forecasting error. To enrich sample size while preserving the marginal distribution and temporal correlation, the thesis utilises the Gaussian kernel density estimate (KDE) to fit the marginals and apply the Gaus-

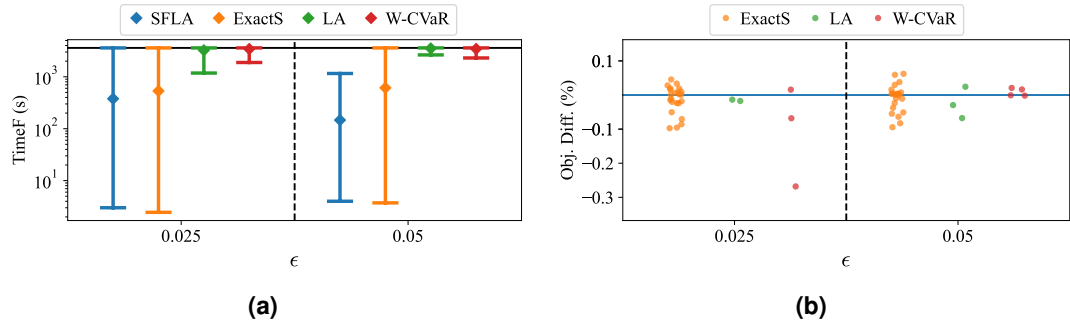


Figure A.2: Performance comparison of the proposed SFLA with benchmarks for the UC problem: ExactS, LA, W-CVaR, and Bonferroni for the UC problem.

The Wasserstein radius θ is set to 0.5 and the optimisation horizon is set to $T = 24$. (a) Comparison of $TimeF$ (s), where dots represent the mean value of the 30 random runs, with error bars indicating the 95% percentile interval (2.5th to 97.5th). The black horizontal line represents the one-hour TimeLimit. (b) Comparison of $Obj. Diff. (\%)$, where each dot represents the result for one of the 30 runs. The blue horizontal line indicates zero difference between the optimal value of a benchmark method compared to the proposed SFLA. Values higher than this horizontal line indicates that the benchmark achieves lower cost (better optimality) than the proposed SFLA. Note that the thesis only keeps certain random runs where the method to be compared and the proposed method are both solved to optimum. Many runs of LA and W-CVaR fail to solve to optimum and thus the plot only shows a few or zero dots for these two methods. The Bonferroni approximation is infeasible due to its over-conservativeness for all runs and is therefore not displayed.

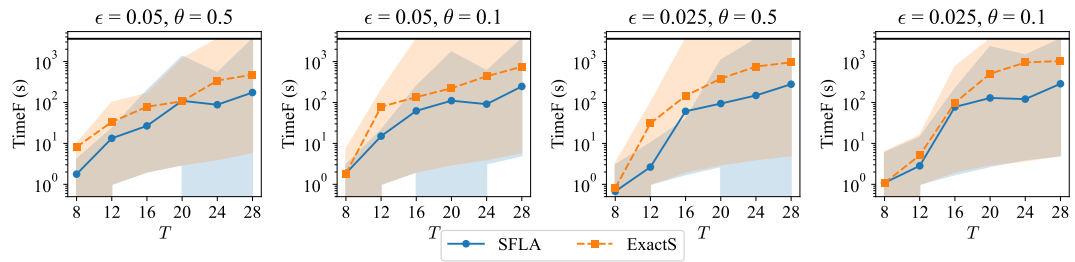


Figure A.3: Computation time to obtain the first comparable high-quality solution under different optimisation horizons T for the proposed SFLA and the benchmark ExactS for the UC problem.

Dots represent the mean value of the 150 random runs, with shaded areas indicating the 95% percentile interval (from 2.5th to 97.5th). The black horizontal line represents the one-hour TimeLimit. The number of historical samples is set to $N = 100$.

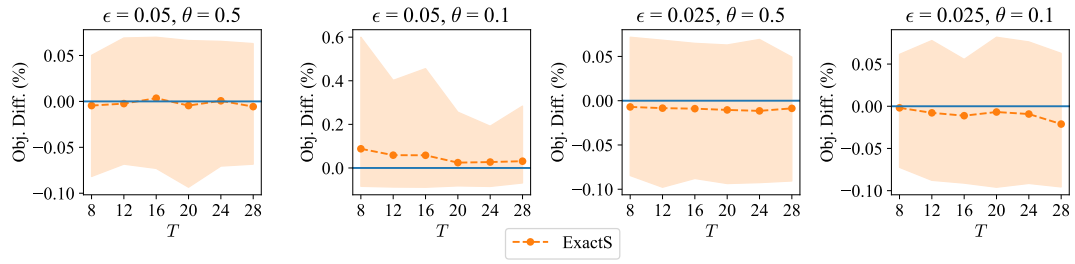


Figure A.4: Comparison of optimality for different optimisation horizons T for the proposed SFLA and the benchmark ExactS for the UC problem.

The round dots represent the mean value of the 150 random runs, with error bars indicating the 95% percentile interval. The blue horizontal line indicates a zero difference of the optimal value of an arbitrary benchmark method compared to the proposed SFLA. Values higher than this horizontal line indicates that the benchmark achieves lower cost (better optimality) than the proposed SFLA. The number of historical samples is set to $N = 100$.

sian copula to generate temporally correlated samples that respect the marginal distribution. Figure A.1 compares the original 156 samples and those generated by the Gaussian KDE plus copula, which show good fitting results that maintain the non-Gaussian marginal and the original temporal correlation.

To ensure robust conclusions, 30 random runs are conducted for each parameter setting of ϵ , θ , N , and T , increasing the number of random runs to 150 for specific evaluations when needed. In each random run, the case study is:

1. Samples a distinct set of generator cost parameters from a uniform distribution based on the range suggested by [77], which impacts the objective function of the UC problem.
2. Samples a different set of historical samples $\{\xi_i\}_{i \in [N]}$ for the random wind forecasting error, uniformly from a total of 1000 training data samples, which affects the feasible region of the UC problem.
3. Selects a random starting simulation hour, uniformly sampled from the 24 hours of a day, which affects the demand and generation profiles and thereby alters the feasible region of the UC problem.
4. Sets a different seed for the optimisation solver, affecting the solver behaviours such as tie-breaking rules.

A.3.2 Supplementary numerical experiments of the unit commitment problem

Figure A.2a compares all methods with an optimisation horizon $T = 24$. As can be seen, in terms of mean values, the proposed SFLA can have 4× faster UC computation than the exact reformulation (ExactS) and even 20× faster than the two linear approximation schemes, namely LA and W-CVaR. This demonstrates the computational efficiency of the proposed SFLA, and also the effectiveness of the strengthening process that makes ExactS even more computationally efficient than linear approximations. Figure A.2b shows that the proposed SFLA has optimality comparable to ExactS and LA, where the minor difference of less than

0.1% is attributed to the preset 0.1% MIPGap. In addition, a random run of W-CVaR shows a 0.25% worse optimality compared to the proposed SFLA, suggesting the higher conservativeness of W-CVaR. The Bonferroni approximation is infeasible for all cases because of its over-conservativeness. Overall, the observations above support both the computational gain and the approximation quality of the proposed SFLA.

Figures A.3 and A.4 compare the proposed SFLA and ExactS under progressively increasing optimisation horizons T , a primary driver of problem complexity. Figure A.3 shows that the proposed SFLA sustains its computational speedup as problem complexity increases. In contrast, the optimality differences illustrated in Figure A.4 exhibit a flat or even decreasing trend for $(\epsilon = 0.05, \theta = 0.1)$. These results indicate that the proposed SFLA preserves computational efficiency in more complex cases without sacrificing optimality, making it particularly advantageous for practical high-complexity applications such as UC.

A.4 Supplementary information for the bilevel bidding problem

A.4.1 Case study settings for the bilevel bidding problem

The case study setting is based on the Alberta Case as described in [151]. The storage efficiencies for both charging and discharging are updated to 95%, in line with the setting in [426], which contrasts to [151] where discharging efficiency was overlooked. The original Alberta Case setup featured 14 generators. Given the complexity of bilevel optimisation and the computational demands of the following intensive numerical simulations, only five generators are selected (G2, G4, G6, G8, and G10). Then the capacities of these generators were scaled proportionally to achieve a combined capacity of 150% of the peak load (8768.89 MW), ensuring sufficient provision of generation and reserve. The modified maximum generation capacities for the five generators are 5248.98 MW, 2470.11 MW, 1482.07 MW, 1914.33 MW and 2037.84 MW respectively. Consistent with the UC case discussed in Section A.3.1, uncertain wind generation is incorporated, which is modelled as a deterministic forecast plus a random forecasting error. The total wind capacity is set to 4800 MW, which accounts for 60% of the total system load and represents a high level of renewable penetration. The wind data follows the UC case in Section A.3.1. The case study examines six time horizons $T \in \{4, 8, 12, 16, 20, 24\}$.

To ensure reliable conclusions, 30 random runs are conducted for each parameter setting of ϵ , θ , and T . In each random run, the case study:

1. Samples a distinct set of generator bid prices from a uniform distribution ranging between 80% and 120% of the bid prices used by [151], which influences the objective function of the bilevel problem.

2. Samples a different set of historical samples $\{\xi_i\}_{i \in [N]}$ for the random wind forecasting error, uniformly from a total of 1000 training data samples, which affects the feasible region of the bilevel problem.
3. Selects a random starting simulation hour, uniformly sampled from the 24 hours of a day, which affects the demand and generation profiles and thereby alters the feasible region of the bilevel problem.
4. Sets a different seed for the optimisation solver, affecting the solver behaviours such as tie-breaking rules.

A.4.2 Supplementary analysis for the bilevel bidding problem

A supplementary analysis of *Bilevel Profit* of the bilevel problem is conducted: among all 720 runs displayed in Figure 3.5, there are only three instances where *Bilevel Profit* of the proposed SFLA is slightly below that of LA and W-CVaR (less than 0.009 k\$). This is caused by another aspect of numerical issue of the proposed SFLA that leads to an incorrect upper bound proved. Setting `NumericalFocus=3` of Gurobi effectively mitigates this issue. Compared to the numerical issue that renders the problem infeasible in LA and W-CVaR, the numerical issue associated with the proposed SFLA is minor.

ϵ	θ	T	<i>Time</i> (s)				<i>TimeF</i> (s)			
			SFLA	LA	W-CVaR	Bonf.	SFLA	LA	W-CVaR	
0.05	0.01	4	6.83	430.86 (63.05×)	294.57 (43.10×)	0.08 (0.01×)	5.63	85.00 (15.09×)	91.67 (16.27×)	
		8	36.84	3200.45 (86.87×)	3332.34 (90.45×)	0.61 (0.02×)	29.63	666.80 (22.50×)	1244.39 (41.99×)	
		12	109.96	3600.03 (32.74×)	3355.00 (30.51×)	– (–)	66.20	1882.83 (28.44×)	2917.42 (44.07×)	
		16	371.21	3600.01 (9.70×)	3379.68 (9.10×)	– (–)	242.21	2916.23 (12.04×)	3600.00 (14.86×)	
		20	669.94	3600.01 (5.37×)	3600.06 (5.37×)	– (–)	562.13	3399.53 (6.05×)	3148.71 (5.60×)	
		24	1256.17	3600.01 (2.87×)	3600.08 (2.87×)	– (–)	933.50	3103.57 (3.32×)	3600.00 (3.86×)	
	0.05	4	7.73	332.43 (42.98×)	265.72 (34.35×)	– (–)	5.93	81.45 (13.73×)	103.09 (17.37×)	
		8	29.54	2595.57 (87.85×)	2906.47 (98.38×)	– (–)	24.97	554.10 (22.19×)	1139.05 (45.62×)	
		12	100.96	3572.06 (35.38×)	3279.78 (32.49×)	– (–)	56.73	2190.86 (38.62×)	2566.10 (45.23×)	
		16	462.03	3600.01 (7.79×)	3588.75 (7.77×)	– (–)	165.63	3240.63 (19.57×)	3542.70 (21.39×)	
		20	1013.09	3600.00 (3.55×)	3486.55 (3.44×)	– (–)	831.77	3461.63 (4.16×)	3203.17 (3.85×)	
		24	1249.38	3600.00 (2.88×)	3600.05 (2.88×)	– (–)	798.47	3519.83 (4.41×)	3386.57 (4.24×)	
0.025	0.01	4	2.55	75.94 (29.74×)	67.70 (26.52×)	0.34 (0.13×)	1.57	25.97 (16.57×)	29.15 (18.61×)	
		8	10.96	744.71 (67.97×)	454.49 (41.48×)	– (–)	8.93	134.24 (15.03×)	1679.56 (188.01×)	
		12	27.53	1633.57 (59.34×)	676.05 (24.56×)	– (–)	25.30	579.79 (22.92×)	305.33 (12.07×)	
		16	44.97	3074.33 (68.37×)	1643.78 (36.55×)	– (–)	39.73	1450.03 (36.49×)	1620.33 (40.78×)	
		20	73.87	3353.31 (45.39×)	1621.46 (21.95×)	– (–)	62.57	2320.12 (37.08×)	3600.00 (57.54×)	
		24	106.10	3573.18 (33.68×)	– (–)	– (–)	88.20	3180.79 (36.06×)	– (–)	
	0.05	4	3.39	39.35 (11.59×)	38.04 (11.21×)	– (–)	2.13	20.11 (9.43×)	17.68 (8.29×)	
		8	10.78	208.10 (19.31×)	139.15 (12.91×)	– (–)	9.53	85.57 (8.98×)	396.08 (41.55×)	
		12	16.75	609.51 (36.38×)	347.29 (20.73×)	– (–)	15.73	214.86 (13.66×)	324.00 (20.59×)	
		16	29.20	1293.95 (44.32×)	965.47 (33.07×)	– (–)	27.33	477.93 (17.49×)	842.50 (30.82×)	
		20	38.53	1960.49 (50.89×)	2035.01 (52.82×)	– (–)	36.27	981.31 (27.06×)	2007.50 (55.35×)	
		24	61.72	2240.78 (36.30×)	2705.43 (43.83×)	– (–)	50.20	1183.65 (23.58×)	2977.83 (59.32×)	

Table A.1: Comparison of the computation time for the bilevel problem

Note. Entries with “–” indicate that no feasible runs were completed for that parameter setting. The *TimeF* (s) of Bonferroni approximation is not displayed because Bonferroni approximation is overly conservative such that it can easily find high *Bilevel Profit* (k\$) (the bilevel objective value) but poor *Actual Profit* (k\$) (plots 3.5d and 3.5l).

Appendix for the Fuel Poverty and Network Impact of Heat Pumps

This appendix provides supplementary information for Chapter 5.

B.1 Key resources table

The key resources used to produce the case study results are listed in Table B.1.

B.2 Resource availability

B.2.1 Data and code availability

The source data is available as listed in the Key Resources Table B.1. Specifically, as an additional output, the distribution network dataset used to support Chapter 5 has been further processed and enriched to create two supplementary datasets. The first is the main dataset for Great Britain's distribution networks, which compiles information on firm capacity, peak demands, locations, and parent transmission nodes (grid supply points, or GSPs) for all primary substations (PSs). The second dataset extends the main network dataset by linking each PS to the number of households using different types of central heating, as recorded in census data (Census 2021 for England and Wales, and Census 2011 for Scotland, since the up-to-date Census 2022 data has not yet been fully released). These datasets are published in [75].

Table B.1: Key Resources Table

RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Scotland household energy consumption, income, heating type, and location	Scottish House Condition Survey [334]	http://doi.org/10.5255/UKDA-SN-8974-1
England household energy consumption, income, heating type, and location	English Housing Survey [44]	http://doi.org/10.5255/UKDA-SN-8891-1
Scotland census data	Scotland's Census 2011 [346]	https://www.scotlandscensus.gov.uk/search-the-census#/
England census data	Census 2021 [340]	https://www.ons.gov.uk/census
Prices of electricity and gas	BEIS quarterly energy prices [337]	https://www.gov.uk/government/statistical-data-sets/annual-domestic-energy-price-statistics
Prices of heating oil and coal	BEIS Consumer Price Index [338]	https://www.gov.uk/government/statistical-data-sets/monthly-domestic-energy-price-statistics
Prices of LPG and biofuels	Nottenergy [339]	https://nottenergy.com/resources/energy-cost-comparison/
Per-unit GHG Emissions by Fuels	UK Government [345]	https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2020
Primary Substation Datasets SP Energy Networks	SPEN [427]	https://www.spenergynetworks.co.uk/pages/long_term_development_statement.aspx
Primary Substation Datasets in Scottish & South Electricity Networks	SSEN [428]	https://network-maps.ssen.co.uk/
Primary Substation Datasets in Western Power Distribution	National Grid [429]	https://www.nationalgrid.co.uk/our-network/network-capacity-map-application
Primary Substation Datasets in UK Power Networks	UKPN [430]	https://ukpowernetworks.opendatasoft.com/explore/dataset/ukpn_primary_postcode_area/information/
Primary Substation Datasets in Electricity North West	ENW [431]	https://www.enwl.co.uk/get-connected/network-information/long-term-development-statement/
Primary Substation Datasets in Northern Power Grid	NPG [350]	https://www.northernpowergrid.com/demand-availability-map
Normalised Heater Demand Profile	[347]	https://data.ukedc.rl.ac.uk/browse/edc/efficiency/residential/Buildings/heat_demand_by_local_area/
Software and algorithm		
Software	Python	https://www.python.org/

Case study outputs

To facilitate subsequent research and provide a better reference, the case study outputs are stored in a CSV file, available in <https://github.com/PSAL0xford/Datasets--Social-Benefits-and-Network-Costs-of-Heat-Pumps>. This file has 41 rows, corresponding to the 32 Scotland LAs and 9 England government regions studied here. The columns include the following:

- fp_2019(%): the estimated fuel poverty rate over the pre-crisis period (Section Fuel poverty estimation).
- fp_2022(%): the estimated fuel poverty rate over the crisis period (Section Fuel poverty estimation).
- fp_2019_HP_plus_upfront(%): the estimated fuel poverty rate over the pre-crisis period after replacing off-gas heating with HPs. The HP upfront cost is included through the amortisation with a 7% interest rate (Section Geographical illustration).
- fp_2022_HP_plus_upfront(%): the estimated fuel poverty rate over the crisis period after replacing off-gas heating with HPs. The HP upfront cost is included through the amortisation with a 7% interest rate (Section Geographical illustration).
- fp_2019_HP_plus_upfront_plus_network(%): the estimated fuel poverty rate over the pre-crisis period after replacing off-gas heating with HPs. The HP upfront cost is included through the amortisation with a 7% interest rate. The additional repayment towards the network upgrade costs is also included (Section Network upgrade cost).
- fp_2022_HP_plus_upfront_plus_network(%): the estimated fuel poverty rate over the crisis period after replacing off-gas heating with HPs. The HP upfront cost is included through the amortisation with a 7% interest rate. The additional repayment towards the network upgrade costs is also included (Section Network upgrade cost).
- off_gas_GHG_emission(kgCO₂e): the total GHG emission by residential heating in the off-gas sector (Section GHG emission reduction).
- off_gas_GHG_emission_HP(kgCO₂e): the total GHG emission by residential heating in the off-gas sector after the HP rollout (Section GHG emission reduction).
- off_gas_GHG_reduction_rate(%): the proportion of reduction of GHG emission by residential heating in the off-gas sector after the HP rollout (Section GHG emission reduction).
- network_upgrade_cost(M£): the total distribution network upgrade cost for a region, scaled from the PS upgrade costs (Section Network upgrade cost).
- network_upgrade_cost_per_PS(M£): the distribution network upgrade cost for a region averaged by the number of PSs in the region (Section Network upgrade cost).
- number_of_PS: the number of PSs in the region.

B.3 Amortisation and levelised cost

Fuel poverty by definition is determined by the annual fuel cost. Therefore, the thesis uses the first-year payment towards the amortisation as this is intuitive and directly reflects the additional payment for each year in Section Upfront cost. However, this is essentially equivalent to the calculation by the levelised lifetime cost method, which is widely used in measuring the per-unit energy cost considering capital investments [344]. This chapter illustrates the equivalence below:

Denote C_{UH} the upfront cost of an HP, E_A the annual HP electricity consumption (kWh), N_{LHP} the expected HP life-span, r_{dis} the annual interest/discount rate. The levelised lifetime cost LC (£/kWh) can be calculated as [344]:

$$LC = \frac{C_{UH} + \sum_{n=1}^{N_{LHP}} \frac{E_A \times \hat{C}_{elec}}{(1+r_{dis})^n}}{\sum_{n=1}^{N_{LHP}} \frac{E_A}{(1+r_{dis})^n}} \quad (B.1)$$

where the maintenance cost is not considered as discussed in Section Estimating fuel poverty. The end-of-life cost is not considered as well for simplicity. Note that LC is usually calculated on an annual basis. Here n starts from 1 to account for the cash flow at the end of each year.

The annual energy bill C_{YU} considering the upfront cost can be derived by scaling up the levelised cost LC , i.e.,

$$\begin{aligned} C_{YU} &= E_A \times LC \\ &= E_A \times \frac{C_{UH} + \sum_{n=1}^{N_{LHP}} \frac{E_A \times \hat{C}_{elec}}{(1+r_{dis})^n}}{\sum_{n=1}^{N_{LHP}} \frac{E_A}{(1+r_{dis})^n}} \\ &= E_A \times \frac{C_{UH}}{\sum_{n=1}^{N_{LHP}} \frac{E_A}{(1+r_{dis})^n}} + E_A \times \hat{C}_{elec} \\ &= \frac{C_{UH}}{\sum_{n=1}^{N_{LHP}} \frac{1}{(1+r_{dis})^n}} + E_A \times \hat{C}_{elec} \\ &= \frac{C_{UH}}{(1 - (\frac{1}{1+r_{dis}})^{N_{LHP}})/r_{dis}} + E_A \times \hat{C}_{elec} \\ &= \frac{C_{UH} r_{dis} (1+r_{dis})^{N_{LHP}}}{(1+r_{dis})^{N_{LHP}} - 1} + E_A \times \hat{C}_{elec} \end{aligned} \quad (B.2)$$

where the second term $E_A \times \hat{C}_{elec}$ is the annual running cost, while the first term $\frac{C_{UH} r_{dis} (1+r_{dis})^{N_{LHP}}}{(1+r_{dis})^{N_{LHP}} - 1}$ is exactly the amount of amortisation paid under an annual cash flow [343]. In many real-world cases the amortisation cash flow is monthly which is used in the calculation of the thesis, while the levelised cost is usually calculated yearly [344]. However, they are essentially equivalent.

B.4 Customers' repayment towards the network cost

Customers will finally repay the network upgrade costs through their increased energy bills [351], which in turn affects the fuel poverty results. To consider this impact, the thesis follows the same setting in [363] to spread the network costs over the 45-year lifespan of upgraded network equipment. In addition, because the future money is worth less, the Net Present Value from the additional customers' bills should equal the total cost of the network upgrade C_{net} for that region:

$$C_{\text{net}} = N_{\text{house}} \sum_{i=0}^{45} \frac{C'_{\text{house}}}{(1 + r_{\text{dis}})^i} \quad (\text{B.3})$$

where C'_{house} is the additional annual energy bill for one household, N_{house} is the total number of households in that region, and r_{dis} is the discount rate. This chapter sets it to 7%, the interest rate under a normal 20-year secured homeowner loan interest rate by a UK loan company [360], which is also used to spread out the HP upfront cost over its life span.

The additional energy bill C'_{house} per household per year can be therefore calculated as follows:

$$C'_{\text{house}} = \frac{C_{\text{net}}}{N_{\text{house}} \sum_{i=0}^{45} \frac{1}{(1 + r_{\text{dis}})^i}} \quad (\text{B.4})$$

Appendix for the Grid Flexibility and Cost of Data Centres

This appendix provides supplementary information and analysis for Chapter 6.

C.1 Cost minimisation problem tightening

The cost minimisation problem (6.15) is a challenging mixed-integer linear programming (MILP). To speed up the computation, the thesis introduces a tightening method by integrating a valid lower bound of the cost into (6.15). This method assumes that there is no queuing time, meaning submission time equals start time.

Based on Eqs. (6.11) and (6.12), the reduction of the computing price for job j can be rewritten as:

$$c_j = A_j \cdot \delta_j \cdot D_j \cdot R \cdot N_j^R = A_j \cdot R \cdot t_j^{\text{Ex}} \cdot N_j^R \quad (\text{C.1})$$

Note that $A_j = A$ has been assumed in Section Scaling the cost of flexibility provision. Under this assumption, the total flexibility cost C^{Flex} becomes:

$$C^{\text{Flex}} = \sum_{j \in \mathcal{J}} c_j = A \left(\sum_{j \in \mathcal{J}} t_j^{\text{Ex}} \cdot N_j^R \right) \cdot R \quad (\text{C.2})$$

Because t_j^{Ex} is the extra time beyond the original job completion time and N_j^R is the number of computing resources, the term $(\sum_{j \in \mathcal{J}} t_j^{\text{Ex}} \cdot N_j^R)$ reflects the effort to complete the “total delayed computing workload of all jobs W^{delay} ”. Furthermore, $(\sum_{j \in \mathcal{J}} t_j^{\text{Ex}} \cdot N_j^R)$ is an upper bound of W^{delay} , because jobs over their delayed times t_j^{Ex} may not 100% utilise their computing resources due to the data centre capacity limit. In other words:

$$W^{\text{delay}} \leq \left(\sum_{j \in \mathcal{J}} t_j^{\text{Ex}} \cdot N_j^R \right) \quad (\text{C.3})$$

when there is no dynamic quota, providing flexibility will definitely trigger workload delays. Based on the linear power model (6.1i), the delayed computing workload is proportional to the shifted energy over flexibility provision through the power coefficient G . Now, consider the provision of S units of flexibility for $|\mathcal{N}^A|$ times with each of the flexibility activation sustaining $|\mathcal{T}_i^A|$ time steps, the total shifted energy is $|\mathcal{T}_i^A| \cdot |\mathcal{N}^A| \cdot S$, so the “total delayed computing workload of all jobs W^{delay} ” can be expressed as:

$$W^{\text{delay}} = |\mathcal{T}_i^A| \cdot |\mathcal{N}^A| \cdot \frac{S}{G} \quad (\text{C.4})$$

which, combined with (C.1) and (C.3), leads to a lower bound $\underline{C}$ of the total flexibility cost C^{Flex} , namely the total reduction in computing prices:

$$\underline{C} = A \cdot |\mathcal{T}_i^A| \cdot |\mathcal{N}^A| \cdot R \cdot \frac{S}{G} \leq C^{\text{Flex}} = \sum_{j \in \mathcal{J}} c_j \quad (\text{C.5})$$

In the simulation of Chapter 6, setting (C.5) as an additional constraint in (6.15) reduces the solution time significantly.

C.2 Cost minimisation problem tightening under dynamic quota

Because job delays could be avoided by using spare resources to accelerate computation, the lower bound in (C.5) is not valid for dynamic quota, and an updated lower bound is required. Suppose S_a is the maximum flexibility without causing any job delays in the dynamic quota opportunity. In case where delivered flexibility $S_b > S_a$, the new valid lower bound for the total computing price reduction is:

$$\underline{C} = A \cdot |\mathcal{T}_i^A| \cdot |\mathcal{N}^A| \cdot R \cdot \frac{S_b - S_a}{G(1 + K^{\text{DQ}})} \quad (\text{C.6})$$

To see this, the profile offering the maximum flexibility S_a under zero job delay can be considered a new baseline profile. Then, providing $(S_b - S_a)$ additional flexibility will certainly trigger job delays and thus the reduction in computing prices. Therefore, the lower bound of the total price reduction can be obtained by substituting S with $(S_b - S_a)$ in (C.5), which is further divided by $1 + K^{\text{DQ}}$ that represents the computing speed-up with dynamic quota.

C.3 Discussions of job partition

The proposed job partition strategy in Section 6.1.4 can lead to conservativeness in finding the maximum flexibility, as jobs completed after the optimisation horizon will have no ability to delay (as in the two examples above). This conservativeness increases for data centres with more long-computing-time jobs, such as the AI-focused HPC data centre, as previously observed in Fig. 6.1c. This conservativeness will reduce with the increase of the optimisation horizon, as illustrated in Fig. C.1. Despite of the conservativeness, an advantage of this partition strategy is the feasibility guarantee for the next-horizon problem, since this problem does not carry out more computing workload than that before flexibility provision. This also means that optimisation problems under non-overlapping horizons are independent, so these problems for different date ranges can be solved in parallel. One may consider another partition strategy such that jobs only need to complete a portion of the computing workload within the current optimisation horizon. However, this may lead to an overestimate of the flexibility cost because some jobs are forced to be delayed unnecessarily.

C.4 Supplementary information

C.4.1 Computing prices and model specifications from cloud platforms

This chapter collected GPU and CPU rental options available on Google Cloud, Amazon Web Services (AWS), Oracle, and Lambda Cloud. These cloud platforms provide price and machine specifications, where the latter enables to link the price information to device power and computing speed from other sources. This section details the source of each price and specification information. Note that on cloud platforms, they usually use the term “instance” to refer to “computing option”.

- For Google Cloud, up-to-date price information and the GPU model can be found at their instance creation page, which requires a Google account log-in. The CPU model specification can be accessed here [432].
- For AWS, computing prices for all computing options can be found here [433]. Model specifications for both the GPU and CPU options are available here [434].
- For Oracle, the computing price and model specifications are available here [435].
- Lambda Cloud provides computing prices and model specifications here [420]. Note that, as mentioned in the main content of Chapter 6, GPU prices on Lambda Cloud are significantly lower than in other cloud platforms. Therefore, this section analysed the Lambda Cloud data separately as Fig. 6.4.

Previous computing price information are for the whole machine. This chapter divides the total price by the number of virtual CPUs (vCPUs) or GPUs to get the price per vCPU or per GPU. This can be justified by the observed proportional relationship between price and number of vCPUs/GPUs on cloud platforms.

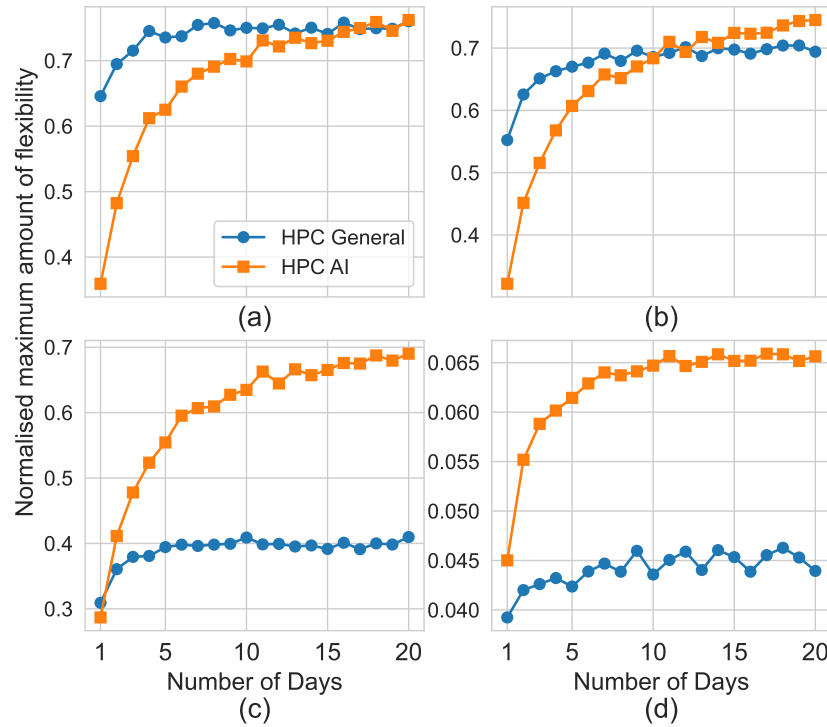


Figure C.1: The maximum amounts of flexibility calculated by solving optimisation problems with varying horizons (number of days).

The calculation is based on the two HPC data centres in Fig. 6.1. (a) Results for a power system service with duration of 0.25 hours and frequency of 365 times/year. (b) Results when the duration is 0.25 hours and the frequency is 2,920 times per year. (c) Results when the duration is 2 hours and the frequency is 365 times per year. (d) Results when the duration is 2 hours and the frequency is 2,920 times per year. The increasing trend of the maximum amount of flexibility with the increasing optimisation horizon confirms the anticipation of reduced conservativeness. This figure shows that the AI-focused HPC data centre in Figs. 6.1c and 6.1d should have maximum flexibility similar to the general-purpose HPC data centre. This holds true even for short-duration, low-frequency services when the optimisation horizon increases beyond 15 days.

On most cloud platforms, CPUs are rented on the basis of vCPUs, which typically represent one thread of a physical CPU. This chapter follows their convention here. Details on the rated power of both CPUs and GPUs are available on the respective manufacturers' websites. The power of a vCPU is calculated by dividing the rated power of the physical CPU by the number of vCPUs.

It should be noted that cloud platforms provide CPUs not just for high-performance computing (HPC) but also for other purposes such as web services. As the thesis focuses on HPC data centres, only a subset of CPU options on cloud platforms with optimised computing performance is collected. The collection criteria are summarised below:

- Google Cloud: For CPUs, the thesis only collects the “compute optimised” instances, and several “general” instances with description “consistently high performance”. Discounts are not applied. The price region is “US-central-1”.

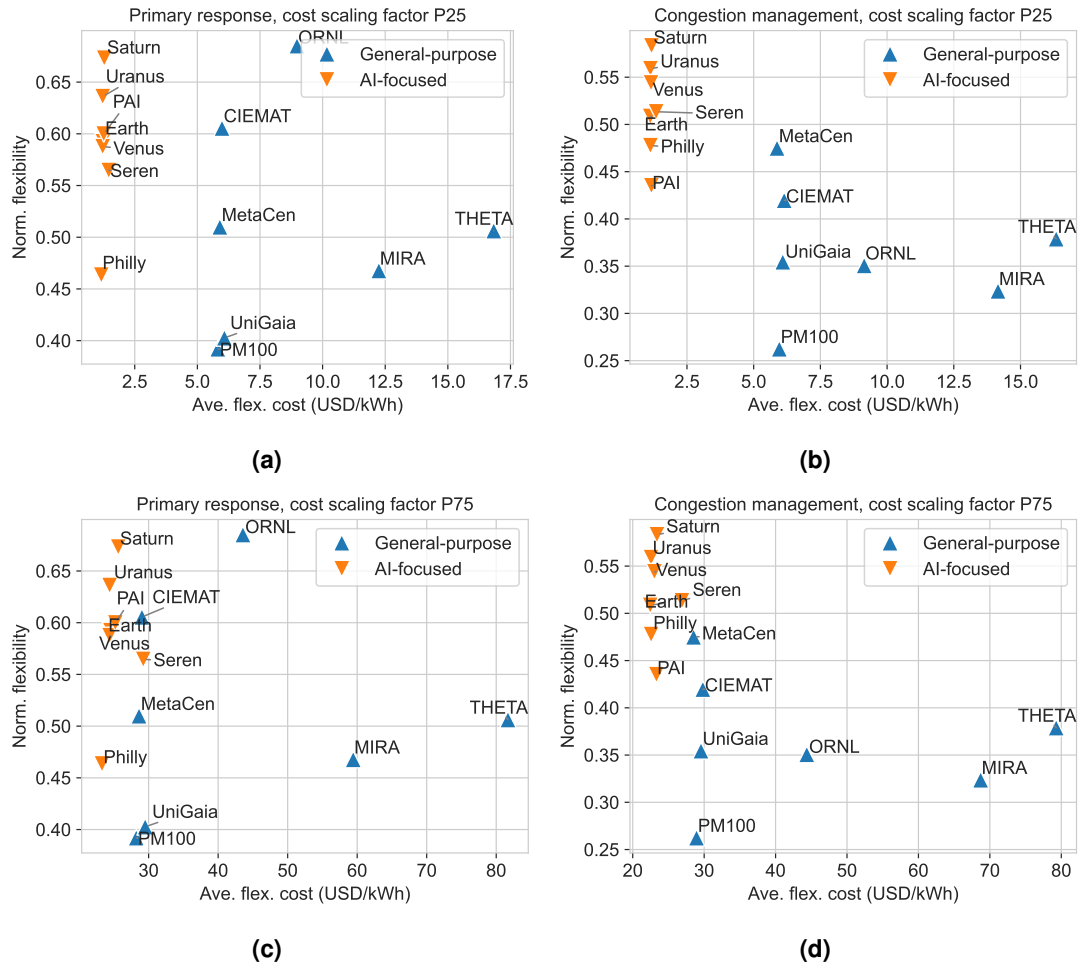


Figure C.2: The normalised maximum amount of flexibility (Norm. flexibility) versus the average flexibility cost when providing 100% of the maximum flexibility (ave. flex. cost) for two power system services.

(a) Results for the primary response service under the 25th percentile (P25) of the cost scaling factor. (b) Results for the congestion management service under the 25th percentile (P25) of the cost scaling factor. (c) Results for the primary response service under the 75th percentile (P75) of the cost scaling factor. (d) Results for the congestion management service under the 75th percentile (P75) of the cost scaling factor. The cost scaling factor is estimated using data from Google Cloud, AWS, and Oracle. **This figure leads to the same conclusion as Fig. 6.5 of the main Chapter 6.**

- AWS: This chapter collects “compute optimised” instances and “HPC optimised” instances. This chapter selects “on-demand” pricing with Linux operation systems. The price zone is “US East (N. Virginia)”.
- Oracle: Oracle does not have dedicated “compute optimised” instances. However, they recommend the “bare metal” type for HPC computing [436]. Therefore, the thesis only selects CPU options within that type.
- Lambda Cloud only provides GPU rental, so is not discussed here.

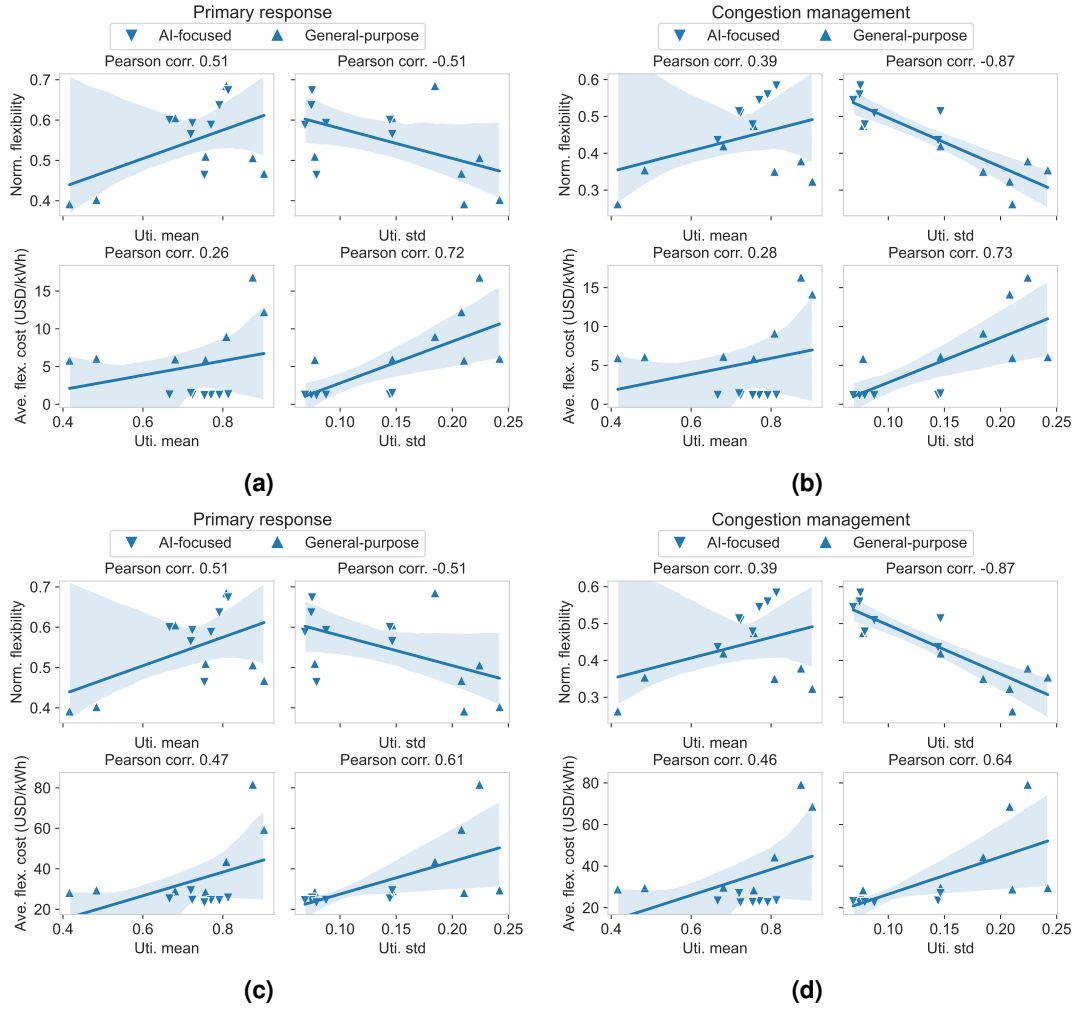


Figure C.3: Correlation analysis of the flexibility results and data centre utilisation patterns.

(a) Correlation between the normalised maximum amount of flexibility (Norm. flexibility) for primary response, the average flexibility cost (Ave. flex. cost) for primary response, the mean utilisation rate (Uti. mean), and the standard deviation of utilisation (Uti. std). Regression lines with confidence intervals (shaded areas) are plotted for highlighting the trends. (b) Correlation results where the flexibility and cost are for congestion management. Plots (a) and (b) are for the 25th percentile of the cost scaling factor. Plots (c) and (d) are for the 75th percentile of the cost scaling factor. The cost scaling factor is estimated using data from Google Cloud, AWS, and Oracle. **This figure leads to the same conclusion, as has been analysed in Fig. 6.6 of the main Chapter 6.**

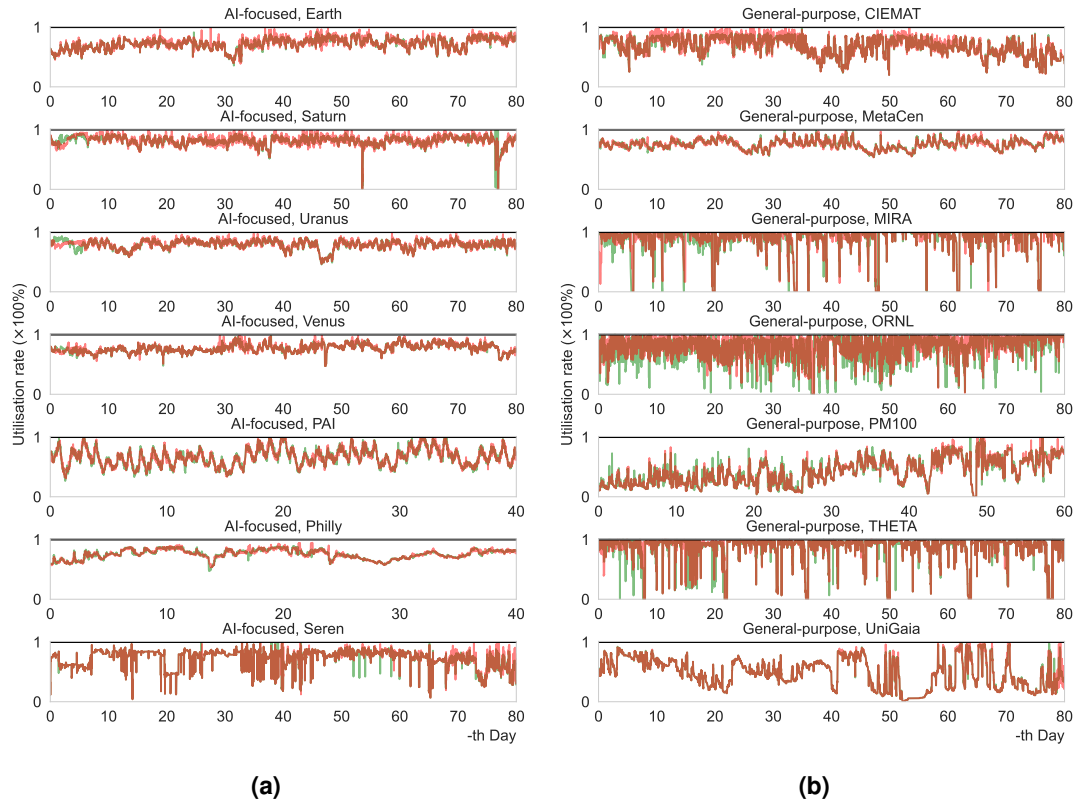


Figure C.4: Baseline utilisation time series of 14 data centres before and after job aggregation.

(a) Utilisation of the 7 AI-focused HPC data centres. (b) Utilisation of the 7 general-purpose HPC data centres. The black lines refer to the utilisation upper bound. The green lines are the utilisation derived by the original job datasets. The red lines are the utilisation derived by the aggregated jobs after job aggregation, which is a method used to reduce the computation complexity of solving optimisation problems, as detailed in the Methods section. It can be seen that the aggregated baseline profiles effectively capture the characteristics of the original profiles before aggregation (the green lines). Therefore, the applied aggregation strategy aims to improve computational efficiency while preserving utilisation fidelity and potential flexibility.

C.4.2 Computing speed information

PassMark CPU benchmark test [393] provides multithread CPU ratings, which are used as the CPU speed. This rating is the weighted harmonic average of the computing speed for several benchmark tests, such as integer math, floating point math, physics, etc., that are important for HPC applications [394].

GPU speed information is available at Lambda Lab [395], which is the weighted average speed for a range of AI tasks. The speed is relative to the speed of a single Tesla V100 GPU, and is provided on both the FP32 and FP16 precision basis. The harmonic average of the two precisions is used as the final GPU speed. The 2023 benchmark speed results are used due to their up-to-dateness, unless only the 2022 speed information is available for a few GPU options.

C.4.3 Data file availability

In order to estimate the cost scaling factor (see Section 6.1.2), the price and speed information from cloud platforms are collected and recorded into the “all_cloud_data_chart.xlsx” file, which is available at Zenodo: <https://doi.org/10.5281/zenodo.14454915>.

C.4.4 Data file description

The “all_cloud_data_chart.xlsx” file has the following columns:

1. **Provider:** Specifies the cloud service provider (e.g., Google Cloud, AWS, Oracle Cloud).
2. **Type:** Indicates whether the entry refers to a CPU (Central Processing Unit) or a GPU (Graphics Processing Unit).
3. **Model:** The specific model of the CPU or GPU used in this computing option.
4. **Number of vCPU / GPU:** Lists the number of vCPUs or GPUs in the computing option. An asterisk (*) indicates that the option is estimated instead of actually available on the cloud platform. For example, a Cloud platform may only rent 4 GPUs or 32 vCPUs as a whole. Then, the price of 1 GPU or 1 vCPU can be estimated, as long as the corresponding speed information is available. The price is estimated by, e.g., dividing the 4 GPU price by 4. This can be justified by the observed proportional relationship between the price and the number of vCPUs/GPUs on the cloud platforms. This is to enrich the dataset to create more accurate estimate; otherwise, there are only limited computing options which can only support limited estimation samples of the cost scaling factor.
5. **Memory (RAM per CPU, VRAM per GPU):** Indicates the total RAM for CPUs or VRAM per GPU.
6. **Unit Price (\$/(unit-h)):** The hourly cost in USD per vCPU or per GPU.

7. **Total device price (\$/h):** The total hourly cost for the entire computing option, considering all vCPUs or GPUs.
8. **CPU Score:** The CPU speed for this option, which is only used for general-purpose HPC data centres.
9. **GPU FP32:** The FP32 GPU speed relative to a single "Tesla V100", which is only used for AI-focused HPC data centres.
10. **GPU FP16:** The FP16 GPU speed relative to a single "Tesla V100", which is only used for AI-focused HPC data centres.
11. **Unit Rated Power (W/(vCPU, GPU)):** The power consumption per vCPU or GPU, measured in watts (W).
12. **Total Rated Power (W):** The total power consumption for all vCPUs or GPUs in the computing option.
13. **Notes:** Provides additional information or clarifications about the data, such as assumptions or special conditions.

Bibliography

- [1] Ofgem, “Distributed Energy: Summary for Non-Specialists,” 2017, accessed: 2025-08-06. [Online]. Available: <https://www.ofgem.gov.uk/sites/default/files/docs/2008/02/de-non-specialist-summary---final.pdf>
- [2] The International Energy Agency, “Unlocking the potential of distributed energy resources,” 2022, accessed: 2025-08-06. [Online]. Available: <https://www.iea.org/reports/unlocking-the-potential-of-distributed-energy-resources>
- [3] National Grid ESO, “Future energy scenarios,” 2024, accessed: 2024-12-10. [Online]. Available: <https://www.neso.energy/publications/future-energy-scenarios-fes>
- [4] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, “Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants,” *Nature energy*, vol. 3, no. 2, pp. 94–101, 2018.
- [5] J. L. Mathieu, G. Verbič, T. Morstyn, M. Almassalkhi, K. Baker, J. Braslavsky, K. Bruninx, Y. Dvorkin, G. S. Ledva, N. Mahdavi *et al.*, “A new definition of demand response in the distributed energy resource era,” *arXiv preprint arXiv:2410.18768*, 2024.
- [6] T. Morstyn and X. Wang, “Opportunities for quantum computing within net-zero power system optimization,” *Joule*, 2024.
- [7] The International Energy Agency, “Digitalisation and energy,” 2017, accessed: 2025-08-06. [Online]. Available: <https://www.iea.org/reports/digitalisation-and-energy>
- [8] C. Crozier, T. Morstyn, and M. McCulloch, “The opportunity for smart charging to mitigate the impact of electric vehicles on transmission and distribution systems,” *Applied Energy*, vol. 268, p. 114973, 2020.
- [9] National Grid ESO, “UK future energy scenarios.” [Online]. Available: <https://www.nationalgrideso.com/future-energy/future-energy-scenarios>
- [10] M. Braun, *Provision of ancillary services by distributed generators: Technological and economic perspective*. kassel university press GmbH, 2009, vol. 10.
- [11] Á. Paredes, J. A. Aguado, C. Essayeh, Y. Xia, I. Savelli, and T. Morstyn, “Stacking revenues from flexible ders in multi-scale markets using tri-level optimization,” *IEEE Transactions on Power Systems*, vol. 39, no. 2, pp. 3949–3961, 2023.
- [12] S. Chatzivasileiadis, “Optimization in modern power systems,” *Lecture Notes. Tech. Univ. of Denmark*. Available online: <https://arxiv.org/pdf/1811.00943.pdf>, 2018.
- [13] A. J. Conejo and L. Baringo, *Power system operations*. Springer, 2018, vol. 11.

- [14] FTI Consulting | Energy Systems Catapult, "Assessment of locational wholesale electricity market design options in GB," FTI Consulting | Energy Systems Catapult, Tech. Rep., October 2023, accessed: 2025-01-01. [Online]. Available: <https://es.catapult.org.uk/report/assessment-of-locational-wholesale-electricity-market-design-options-in-gb/>
- [15] Y. Fu, M. Shahidehpour, and Z. Li, "Security-constrained unit commitment with ac constraints," *IEEE transactions on power systems*, vol. 20, no. 2, pp. 1001–1013, 2005.
- [16] Y. Fu, Z. Li, and L. Wu, "Modeling and solution of the large-scale security-constrained unit commitment," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 3524–3533, 2013.
- [17] G. Byeon and P. Van Hentenryck, "Unit commitment with gas network awareness," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1327–1339, 2019.
- [18] H.-D. Chiang, Z.-Y. Wang, and L. Zeng, "Dynamic relationship between kkt saddle solutions and optimal solutions in ac opf problems," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1637–1646, 2023.
- [19] K. Baker, "Solutions of DC OPF are never AC feasible," in *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, 2021, pp. 264–268.
- [20] L. Piloto, S. Liguori, S. Madjiheurem, M. Zgubic, S. Lovett, H. Tomlinson, S. Elster, C. Apps, and S. Witherspoon, "Canos: A fast and scalable neural ac-opf solver robust to n-1 perturbations," *arXiv preprint arXiv:2403.17660*, 2024.
- [21] Y. Chen, A. Casto, F. Wang, Q. Wang, X. Wang, and J. Wan, "Improving large scale day-ahead security constrained unit commitment performance," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4732–4743, 2016.
- [22] X. Sun, P. B. Luh, M. A. Bragin, Y. Chen, J. Wan, and F. Wang, "A novel decomposition and coordination approach for large day-ahead unit commitment with combined cycle units," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5297–5308, 2018.
- [23] Y. Fu, Z. Li, and L. Wu, "Modeling and solution of the large-scale security-constrained unit commitment," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 3524–3533, 2013.
- [24] S. Riaz and P. Mancarella, "Modelling and characterisation of flexibility from distributed energy resources," *IEEE Transactions on Power Systems*, 2021.
- [25] X. Chen, E. Dall'Anese, C. Zhao, and N. Li, "Aggregate power flexibility in unbalanced distribution systems," *IEEE Transactions on Smart Grid*, 2020.
- [26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [27] Z. Chen, C. Guo, S. Dong, Y. Ding, and H. Mao, "Distributed robust dynamic economic dispatch of integrated transmission and distribution systems," *IEEE Transactions on Industry Applications*, vol. 57, no. 5, pp. 4500–4512, 2021.

- [28] J. Zhai *et al.*, “Distributionally robust joint chance-constrained dispatch for integrated transmission-distribution systems via distributed optimization,” *IEEE Transactions on Smart Grid*, 2022.
- [29] Y. Wen, Z. Hu, S. You, and X. Duan, “Aggregate feasible region of ders: Exact formulation and approximate models,” *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4405–4423, 2022.
- [30] Z. Tan, Z. Yan, H. Zhong, and Q. Xia, “Non-iterative solution for coordinated optimal dispatch via equivalent projection—part i: Theory,” *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 890–898, 2024.
- [31] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, “Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond,” pp. 896–913, 2016.
- [32] H. Zhou, Y. Zhou, H. Zhang, J. Hu, L. Nordströmd, and G. Yang, “Lstm network-based method for flexibility prediction of aggregated electric vehicles in smart grid,” in *Proceedings of 2020 International Top-Level Forum on Engineering Science and Technology Development Strategy and The 5th PURPLE MOUNTAIN FORUM (PMF2020)*. Springer, 2021, pp. 962–974.
- [33] J. Hu, H. Zhou, Y. Zhou, H. Zhang, L. Nordströmd, and G. Yang, “Flexibility prediction of aggregated electric vehicles and domestic hot water systems in smart grids,” *Engineering*, vol. 7, no. 8, pp. 1101–1114, 2021.
- [34] S. Kwon, “Qos-aware data center operations based on chance-and risk-constrained optimization,” *IEEE Transactions on Cloud Computing*, vol. 10, no. 4, pp. 2887–2896, 2020.
- [35] D. S. Kirschen and G. Strbac, *Fundamentals of power system economics*. John Wiley & Sons, 2018.
- [36] B.-M. Hodge, A. Florita, K. Orwig, D. Lew, and M. Milligan, “Comparison of wind power and load forecasting error distributions,” National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2012.
- [37] A. Prékopa, *Stochastic programming*. Springer Science & Business Media, 2013, vol. 324.
- [38] Y. Ding, T. Morstyn, and M. D. McCulloch, “Distributionally robust joint chance-constrained optimization for networked microgrids considering contingencies and renewable uncertainty,” *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 2467–2478, 2022.
- [39] D. Kuhn, S. Shafiee, and W. Wiesemann, “Distributionally robust optimization,” 2024. [Online]. Available: <https://arxiv.org/abs/2411.02549>
- [40] legislation.gov.uk, “Fuel poverty (targets, definition and strategy) (scotland) act 2019,” 2019. [Online]. Available: <https://www.legislation.gov.uk/asp/2019/10>

- [41] Marmot Review Team for Friends of the Earth, "The health impacts of cold homes and fuel poverty," 2011. [Online]. Available: <https://www.instituteofhealthequity.org/resources-reports/the-health-impacts-of-cold-homes-and-fuel-poverty>
- [42] S. Hinson, P. Bolton, and S. Kennedy, "Fuel poverty in the UK," 2023. [Online]. Available: <https://commonslibrary.parliament.uk/research-briefings/cbp-8730/>
- [43] SHCS Project Team, "Scottish house condition survey: Methodology notes 2019," Scottish Government, Tech. Rep., January 2021. [Online]. Available: <https://www.gov.scot/publications/scottish-house-condition-survey-2019-key-findings/pages/6/>
- [44] Department for Business, Energy & Industrial Strategy. (2022), "English housing survey: Fuel poverty dataset, 2019," UK Data Service, 2022.
- [45] Y. Guan, J. Yan, Y. Shan, Y. Zhou, Y. Hang, R. Li, Y. Liu, B. Liu, Q. Nie, B. Bruckner *et al.*, "Burden of the global energy price crisis on households," *Nature Energy*, vol. 8, pp. 1–13, 2023.
- [46] P. K. Ozili, "Global economic consequence of russian invasion of ukraine," *Available at SSRN*, 2022.
- [47] S. Evans, "Analysis: why UK energy bills are soaring to record highs—and how to cut them," 2022, accessed: 2023-04-12. [Online]. Available: <https://www.carbonbrief.org/analysis-why-uk-energy-bills-are-soaring-to-record-highs-and-how-to-cut-them/>
- [48] B. K. Sovacool, P. Upham, M. Martiskainen, K. E. Jenkins, G. A. Torres Contreras, and N. Simcock, "Policy prescriptions to address energy and transport poverty in the united kingdom," *Nature Energy*, vol. 8, pp. 1–11, 2023.
- [49] Y. Zhou, C. Essayeh, S. Darby, and T. Morstyn, "Evaluating the social benefits and network costs of heat-pumps as an energy crisis intervention," *iScience*, 2024.
- [50] National Energy System Operator (NESO), "Historic GB Generation Mix," accessed: 2025-08-09. [Online]. Available: https://www.neso.energy/data-portal/historic-generation-mix/historic_gb_generation_mix
- [51] Institute for Government, "Electricity market explainer," 2023, accessed: Nov. 01, 2023. [Online]. Available: <https://www.instituteforgovernment.org.uk/article/explainer/electricity-market>
- [52] T. Savage, J. Akroyd, S. Mosbach, M. Hillman, F. Sielker, and M. Kraft, "Universal digital twin—the impact of heat pumps on social inequality," *Advances in Applied Energy*, vol. 5, p. 100079, 2022.
- [53] J. Barnes and S. M. Bhagavathy, "The economics of heat pumps and the (un) intended consequences of government policy," *Energy Policy*, vol. 138, p. 111198, 2020.
- [54] Department for Energy Security and Net Zero, "Fuel poverty statistics," 2020. [Online]. Available: <https://www.gov.uk/government/collections/fuel-poverty-statistics>
- [55] House of Commons Library, "Constituency data: Households off the gas grid," <https://commonslibrary.parliament.uk/constituency-data-households-off-the-gas-grid/>, 2025.

- [56] DESNZ, "Desnz public attitudes tracker: Heat and energy in the home winter 2023, UK," https://assets.publishing.service.gov.uk/media/65fc5d5e65ca2f001b7da82c/DESNZ_Public_Attitudes_Tracker_Winter_2023_Heat_and_Energy_in_the_Home__Revised_.pdf, 2024, accessed: 2025-08-08.
- [57] J. Love, A. Z. Smith, S. Watson, E. Oikonomou, A. Summerfield, C. Gleeson, P. Biddulph, L. F. Chiu, J. Wingfield, C. Martin, A. Stone, and R. Lowe, "The addition of heat pump electricity load profiles to gb electricity demand: Evidence from a heat pump field trial," *Applied Energy*, vol. 204, pp. 332–342, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261917308954>
- [58] C. Crozier, "The impact of domestic electric vehicle charging on electricity networks," Ph.D. dissertation, University of Oxford, 2019.
- [59] M. Arregoces and M. Portolani, *Data center fundamentals*. Cisco Press, 2003.
- [60] T. Sterling, M. Brodowicz, and M. Anderson, *High performance computing: modern systems and practices*. Morgan Kaufmann, 2017.
- [61] B. Liu, D. Zydek, H. Selvaraj, and L. Gewali, "Accelerating high performance computing applications: Using cpus, gpus, hybrid cpu/gpu, and fpgas," in *2012 13th International Conference on Parallel and Distributed Computing, Applications and Technologies*. IEEE, 2012, pp. 337–342.
- [62] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," 2021.
- [63] World Economic Forum, "How venture capital is investing in AI in the top five global economies — and shaping the AI ecosystem," <https://www.weforum.org/agenda/2024/05/these-5-countries-are-leading-the-global-ai-race-heres-how-theyre-doing-it/>, [Accessed 15-07-2024].
- [64] Electric Power Research Institute (EPRI), "Powering intelligence: Analyzing artificial intelligence and data center energy consumption," , Tech. Rep., 2024. [Online]. Available: https://www.wpr.org/wp-content/uploads/2024/06/3002028905_Powering-Intelligence_-Analyzing-Artificial-Intelligence-and-Data-Center-Energy-Consumption.pdf
- [65] D. Mytton, M. Ashtine, S. Wheeler, and D. Wallom, "Stretched grid? Managing data center energy demand and grid capacity," *Oxford Open Energy*, vol. 2, no. October, pp. 1–4, feb 2023.
- [66] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier, and J. Dean, "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, 2022.
- [67] "NVIDIA Blackwell Platform Arrives to Power a New Era of Computing," <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>, [Accessed 11-06-2024].

- [68] S. Sorrell, "Jevons' paradox revisited: The evidence for backfire from improved energy efficiency," *Energy Policy*, vol. 37, no. 4, pp. 1456–1469, 2009. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0301421508007428>
- [69] I. Alaperä, S. Honkapuro, and J. Paananen, "Data centers as a source of dynamic flexibility in smart grids," *Appl. Energy*, vol. 229, no. April, pp. 69–79, 2018.
- [70] K. Kaur, S. Garg, N. Kumar, G. S. Aujla, K. K. R. Choo, and M. S. Obaidat, "An Adaptive Grid Frequency Support Mechanism for Energy Management in Cloud Data Centers," *IEEE Syst. J.*, vol. 14, no. 1, pp. 1195–1205, 2020.
- [71] Q. Hu, P. Sun, S. Yan, Y. Wen, and T. Zhang, "Characterization and prediction of deep learning workloads in large-scale gpu datacenters," in *SC:21*, 2021, pp. 1–15.
- [72] Y. Zhou, Y. Xia, H. Yang, and T. Morstyn, "Strengthened and faster linear approximation to joint chance constraints with wasserstein ambiguity," *arXiv preprint arXiv:2412.12992*, 2024.
- [73] Y. Zhou, C. Essayeh, and T. Morstyn, "Aggregated feasible active power region for distributed energy resources with a distributionally robust joint probabilistic guarantee," *IEEE Transactions on Power Systems*, vol. 40, no. 1, pp. 556–571, 2025.
- [74] —, "A novel surrogate polytope method for day-ahead virtual power plant scheduling with joint probabilistic constraints," *Electric Power Systems Research*, vol. 234, p. 110623, 2024.
- [75] —, "Datasets of great britain primary substations integrated with household heating information," *Data in Brief*, vol. 54, p. 110483, 2024.
- [76] Y. Zhou, A. Paredes, C. Essayeh, and T. Morstyn, "Ai-focused hpc data centers can provide more power grid flexibility and at lower cost," *arXiv preprint arXiv:2410.17435*, 2024.
- [77] T. Xu, A. B. Birchfield, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Application of large-scale synthetic power system models for energy economic studies," 2017.
- [78] F. Li and R. Bo, "Dcopf-based Imp simulation: Algorithm, comparison with acopf, and sensitivity," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1475–1485, 2007.
- [79] K. Purchala, L. Meeus, D. Van Dommelen, and R. Belmans, "Usefulness of DC power flow for active power flow analysis," in *IEEE power engineering society general meeting, 2005*. IEEE, 2005, pp. 454–459.
- [80] Y. Fu and M. Shahidepour, "Fast scuc for large-scale power systems," *IEEE Transactions on power systems*, vol. 22, no. 4, pp. 2144–2151, 2007.
- [81] J. Hou, Q. Zhai, Y. Zhou, and X. Guan, "A fast solution method for large-scale unit commitment based on lagrangian relaxation and dynamic programming," *IEEE Transactions on Power Systems*, vol. 39, no. 2, pp. 3130–3140, 2023.
- [82] M. Ma, L. Fan, and Z. Miao, "Consensus admm and proximal admm for economic dispatch and ac opf with socp relaxation," in *2016 North American power symposium (NAPS)*. IEEE, 2016, pp. 1–6.

- [83] A. Castillo, C. Laird, C. A. Silva-Monroy, J.-P. Watson, and R. P. O'Neill, "The unit commitment problem with AC optimal power flow constraints," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4853–4866, 2016.
- [84] T. Wang and H.-D. Chiang, "Theoretical study of non-iterative holomorphic embedding methods for solving nonlinear power flow equations: Algebraic property," *IEEE Transactions on Power Systems*, vol. 36, no. 4, pp. 2934–2945, 2020.
- [85] F. Zohrizadeh, C. Jozs, M. Jin, R. Madani, J. Lavaei, and S. Sojoudi, "A survey on conic relaxations of optimal power flow problem," *European journal of operational research*, vol. 287, no. 2, pp. 391–409, 2020.
- [86] M. He, Z. Soltani, M. Ghaljehei, M. Esmaili, S. Ma, M. Chen, M. Khorsand, R. Ayyanar, and V. Vittal, "A socp-based acopf for operational scheduling of three-phase unbalanced distribution systems and coordination of pv smart inverters," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 229–244, 2023.
- [87] A. Mohammadi, M. Mehrtash, and A. Kargarian, "Diagonal quadratic approximation for decentralized collaborative tso+ dso optimal power flow," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2358–2370, 2018.
- [88] M. H. Ullah and J.-D. Park, "Transactive energy market operation through coordinated tso-dsos-ders interactions," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1978–1990, 2022.
- [89] Z. Yuan and M. R. Hesamzadeh, "Hierarchical coordination of tso-dso economic dispatch considering large-scale integration of distributed energy resources," *Applied energy*, vol. 195, pp. 600–615, 2017.
- [90] —, "A distributed economic dispatch mechanism to implement distribution locational marginal pricing," in *2018 Power Systems Computation Conference (PSCC)*. IEEE, 2018, pp. 1–7.
- [91] N. Kokash, "An introduction to heuristic algorithms," *Department of Informatics and Telecommunications*, vol. 1, pp. 1–7, 2005.
- [92] J. Soares, M. Silva, T. Sousa, Z. Vale, and H. Morais, "Distributed energy resource short-term scheduling using signaled particle swarm optimization," *Energy*, vol. 42, no. 1, pp. 466–476, 2012.
- [93] F. Charbonnier, B. Peng, J. Vienne, E. Stai, T. Morstyn, and M. McCulloch, "Centralised rehearsal of decentralised cooperation: Multi-agent reinforcement learning for the scalable coordination of residential energy flexibility," *Applied Energy*, vol. 377, p. 124406, 2025.
- [94] S. Wang, W. Wu, Q. Chen, J. Yu, and P. Wang, "Stochastic flexibility evaluation for virtual power plant by aggregating distributed energy resources," *CSEE Journal of Power and Energy Systems*, 2022.
- [95] S. Wang and W. Wu, "Aggregate flexibility of virtual power plants with temporal coupling constraints," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5043–5051, 2021.

- [96] J. Zhen and D. Den Hertog, "Computing the maximum volume inscribed ellipsoid of a polytopic projection," *INFORMS Journal on Computing*, vol. 30, no. 1, pp. 31–42, 2018.
- [97] S. Barot and J. A. Taylor, "A concise, approximate representation of a collection of loads described by polytopes," *International Journal of Electrical Power & Energy Systems*, vol. 84, pp. 55–63, 2017.
- [98] M. S. Nazir *et al.*, "Inner approximation of minkowski sums: A union-based approach and applications to aggregated energy resources," in *2018 IEEE Conference on Decision and Control*, 2018, pp. 5708–5715.
- [99] G. Papazoglou and P. Biskas, "Review of methodologies for the assessment of feasible operating regions at the TSO–DSO interface," *Energies*, vol. 15, no. 14, p. 5147, 2022.
- [100] X. Chen and N. Li, "Leveraging two-stage adaptive robust optimization for power flexibility aggregation," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 3954–3965, 2021.
- [101] K. Girigoudar and L. A. Roald, "Identifying secure operating ranges for der control using bilevel optimization," *IEEE Transactions on Smart Grid*, vol. 15, no. 3, pp. 2921–2933, 2023.
- [102] Y. Wen, Z. Hu, J. He, and Y. Guo, "Improved inner approximation for aggregating power flexibility in active distribution networks and its applications," *IEEE Transactions on Smart Grid*, vol. 15, no. 4, pp. 3653–3665, 2023.
- [103] A. Shapiro and A. Philpott, "A tutorial on stochastic programming," *Manuscript. Available at www2.isye.gatech.edu/ashapiro/publications.html*, vol. 17, 2007.
- [104] W. Wei, "Tutorials on advanced optimization methods," *arXiv preprint arXiv:2007.13545*, 2020.
- [105] A. Prékopa, *Stochastic programming*. Springer Science & Business Media, 2013, vol. 324.
- [106] X. Geng and L. Xie, "Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization," *Annual reviews in control*, vol. 47, 2019.
- [107] Z. Chen, D. Kuhn, and W. Wiesemann, "Data-driven chance constrained programs over wasserstein balls," *Operations Research*, vol. 72, no. 1, pp. 410–424, 2024.
- [108] N. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz, and D. Lee, "Distributionally robust chance-constrained programs with right-hand side uncertainty under wasserstein ambiguity," *Mathematical Programming*, 2021.
- [109] J. Luedtke and S. Ahmed, "A sample approximation approach for optimization with probabilistic constraints," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 674–699, 2008.
- [110] M. C. Campi and S. Garatti, "The exact feasibility of randomized solutions of uncertain convex programs," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1211–1230, 2008.

- [111] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.
- [112] H. E. Scarf, K. Arrow, and S. Karlin, *A min-max solution of an inventory problem*. Rand Corporation Santa Monica, 1957.
- [113] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.
- [114] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *Mathematics of Operations Research*, vol. 48, no. 2, pp. 603–655, 2023.
- [115] G. C. Calafiore and L. E. Ghaoui, "On distributionally robust chance-constrained linear programs," *Journal of Optimization Theory and Applications*, vol. 130, pp. 1–22, 2006.
- [116] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, "Ambiguous joint chance constraints under mean and dispersion information," *Operations Research*, vol. 65, no. 3, pp. 751–767, 2017.
- [117] N. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz, and D. Lee, "Strong formulations for distributionally robust chance-constrained programs with left-hand side uncertainty under wasserstein ambiguity," *INFORMS Journal on Optimization*, vol. 5, no. 2, pp. 211–232, 2023.
- [118] W. Xie, "On distributionally robust chance constrained programs with wasserstein distance," *Mathematical Programming*, vol. 186, no. 1, pp. 115–155, 2021.
- [119] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2024. [Online]. Available: <https://www.gurobi.com>
- [120] N. Jiang and W. Xie, "Also-x and also-x+: Better convex approximations for chance constrained programs," *Operations Research*, vol. 70, no. 6, pp. 3581–3600, 2022.
- [121] —, "Also-x#: Better convex approximations for distributionally robust chance constrained programs," *Mathematical Programming*, pp. 1–64, 2024.
- [122] Z. Chen, D. Kuhn, and W. Wiesemann, "On approximations of data-driven chance constrained programs over wasserstein balls," *Operations Research Letters*, vol. 51, no. 3, pp. 226–233, 2023.
- [123] F. Qiu and J. Wang, "Chance-constrained transmission switching with guaranteed wind power utilization," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1270–1278, 2015.
- [124] Z. Liu, Q. Wu, S. S. Oren, S. Huang, R. Li, and L. Cheng, "Distribution locational marginal pricing for optimal electric vehicle charging through chance constrained mixed-integer programming," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 644–654, 2016.

- [125] C. Ning and F. You, "Deep learning based distributionally robust joint chance constrained economic dispatch under wind power uncertainty," *IEEE Transactions on Power Systems*, vol. 37, no. 1, pp. 191–203, 2021.
- [126] Y. Yang, W. Wu, B. Wang, and M. Li, "Analytical reformulation for stochastic unit commitment considering wind power uncertainty with gaussian mixture model," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2769–2782, 2020.
- [127] B.-M. Hodge *et al.*, "Wind power forecasting error distributions: An international comparison," NREL, Tech. Rep., 2012.
- [128] Z.-S. Zhang, Y.-Z. Sun, D. W. Gao, J. Lin, and L. Cheng, "A versatile probability distribution model for wind power forecast errors and its application in economic dispatch," *IEEE transactions on power systems*, vol. 28, no. 3, pp. 3114–3125, 2013.
- [129] C. Tang, J. Xu, Y. Sun, J. Liu, X. Li, D. Ke, J. Yang, and X. Peng, "Look-ahead economic dispatch with adjustable confidence interval based on a truncated versatile distribution model for wind power," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 1755–1767, 2017.
- [130] Z. Wang, C. Shen, F. Liu, X. Wu, C.-C. Liu, and F. Gao, "Chance-constrained economic dispatch with non-gaussian correlated wind power uncertainty," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4880–4893, 2017.
- [131] B. Khorramdel, A. Zare, C. Chung, and P. Gavriladis, "A generic convex model for a chance-constrained look-ahead economic dispatch problem incorporating an efficient wind power distribution modeling," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 873–886, 2019.
- [132] W. van Ackooij, E. C. Finardi, and G. M. Ramalho, "An exact solution method for the hydrothermal unit commitment under wind power uncertainty with joint probability constraints," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6487–6500, 2018.
- [133] A. Bagheri, C. Zhao, and Y. Guo, "Data-driven chance-constrained stochastic unit commitment under wind power uncertainty," in *2017 IEEE Power & Energy Society General Meeting*. IEEE, 2017, pp. 1–5.
- [134] W.-S. Tan and M. Shaaban, "A hybrid stochastic/deterministic unit commitment based on projected disjunctive milp reformulation," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 5200–5201, 2016.
- [135] Q. Wang, Y. Guan, and J. Wang, "A chance-constrained two-stage stochastic program for unit commitment with uncertain wind power output," *IEEE transactions on power systems*, vol. 27, no. 1, pp. 206–215, 2011.
- [136] C. Zhao, Q. Wang, J. Wang, and Y. Guan, "Expected value and chance constrained stochastic unit commitment ensuring wind power utilization," *IEEE Transactions on Power Systems*, vol. 29, no. 6, pp. 2696–2705, 2014.

- [137] M. S. Modarresi, L. Xie, M. C. Campi, S. Garatti, A. Care, A. A. Thatte, and P. Kumar, "Scenario-based economic dispatch with tunable risk levels in high-renewable power systems," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 5103–5114, 2018.
- [138] H. Ming, L. Xie, M. C. Campi, S. Garatti, and P. Kumar, "Scenario-based economic dispatch with uncertain demand response," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1858–1868, 2017.
- [139] Y. Wang, Y. Liu, and D. S. Kirschen, "Scenario reduction with submodular optimization," *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 2479–2480, 2016.
- [140] X. Geng and L. Xie, "Chance-constrained unit commitment via the scenario approach," in *2019 North American Power Symposium (NAPS)*. IEEE, 2019, pp. 1–6.
- [141] Y. Zhang, S. Shen, and J. L. Mathieu, "Distributionally robust chance-constrained optimal power flow with uncertain renewables and uncertain reserves provided by loads," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1378–1388, 2016.
- [142] Z. Shi, H. Liang, S. Huang, and V. Dinavahi, "Distributionally robust chance-constrained energy management for islanded microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2234–2244, 2018.
- [143] W. Xie and S. Ahmed, "Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 1860–1867, 2017.
- [144] A. Arrigo, C. Ordoudis, J. Kazempour, Z. De Grève, J.-F. Toubéau, and F. Vallée, "Wasserstein distributionally robust chance-constrained optimization for energy and reserve dispatch: An exact and physically-bounded formulation," *European Journal of Operational Research*, vol. 296, no. 1, pp. 304–322, 2022.
- [145] C. Ordoudis, V. A. Nguyen, D. Kuhn, and P. Pinson, "Energy and reserve dispatch with distributionally robust joint chance constraints," *Operations Research Letters*, vol. 49, no. 3, pp. 291–299, 2021.
- [146] J. Zhai, Y. Jiang, Y. Shi, C. N. Jones, and X.-P. Zhang, "Distributionally robust joint chance-constrained dispatch for integrated transmission-distribution systems via distributed optimization," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 2132–2147, 2022.
- [147] Y. Wen, Y. Guo, Z. Hu, and G. Hug, "Multiple joint chance constraints approximation for uncertainty modeling in dispatch problems," *IEEE Transactions on Power Systems*, pp. 1–14, 2024.
- [148] L. A. Roald, D. Pozo, A. Papavasiliou, D. K. Molzahn, J. Kazempour, and A. Conejo, "Power systems optimization under uncertainty: A review of methods and applications," *Electric Power Systems Research*, vol. 214, p. 108725, 2023.

- [149] C. N. Dimitriadis, E. G. Tsimopoulos, and M. C. Georgiadis, "Strategic bidding of an energy storage agent in a joint energy and reserve market under stochastic generation," *Energy*, vol. 242, p. 123026, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360544221032758>
- [150] Y. Jia, Z. Mi, Y. Yu, Z. Song, and C. Sun, "A bilevel model for optimal bidding and offering of flexible load aggregator in day-ahead energy and reserve markets," *IEEE Access*, vol. 6, pp. 67 799–67 808, 2018.
- [151] E. Nasrolahpour, J. Kazempour, H. Zareipour, and W. D. Rosehart, "A bilevel model for participation of a storage system in energy and reserve markets," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 582–598, 2017.
- [152] Y. Xia, I. Savelli, and T. Morstyn, "An incentive regulation approach for balancing stakeholder interests in transmission merchant investment," *Electric Power Systems Research*, vol. 234, p. 110617, 2024.
- [153] —, "Integrating local market operations into transmission investment: A tri-level optimization approach," *Applied Energy*, vol. 378, p. 124721, 2025.
- [154] —, "Scalable multi-level optimization for sequentially cleared energy markets with a case study on gas and carbon aware unit commitment," *arXiv preprint arXiv:2502.13643*, 2025.
- [155] A. De Paola, I. Savelli, and T. Morstyn, "A novel ex-ante tariff scheme for cost recovery of transmission investments under elasticity of demand," in *2020 17th International Conference on the European Energy Market (EEM)*. IEEE, 2020, pp. 1–6.
- [156] D. Qiu, Z. Dong, G. Ruan, H. Zhong, G. Strbac, and C. Kang, "Strategic retail pricing and demand bidding of retailers in electricity market: A data-driven chance-constrained programming," *Advances in Applied Energy*, vol. 7, p. 100100, 2022.
- [157] P. A. Gade, H. W. Bindner, and J. Kazempour, "Leveraging p90 requirement: Flexible resources bidding in nordic ancillary service markets," in *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2024, pp. 505–510.
- [158] D. M. Gonzalez *et al.*, "Determination of the time-dependent flexibility of active distribution networks to control their tso-dso interconnection power flow," in *Power Systems Computation Conference (PSCC)*, 2018.
- [159] N. Savvopoulos, N. Hatziaargyriou, and H. Laaksonen, "A holistic approach to the efficient estimation of operational flexibility from distributed resources," *IEEE Open Access Journal of Power and Energy*, 2024.
- [160] S. Taheri, V. Kekatos, H. Veeramachaneni, and B. Zhang, "Data-driven modeling of aggregate flexibility under uncertain and non-convex device models," *IEEE Transactions on Smart Grid*, pp. 1–1, 2022.

- [161] Z. Li, J. Wang, H. Sun, F. Qiu, and Q. Guo, "Robust estimation of reactive power for an active distribution system," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3395–3407, 2019.
- [162] Z. Tan, H. Zhong, Q. Xia, C. Kang, X. S. Wang, and H. Tang, "Estimating the robust p-q capability of a technical virtual power plant under uncertainties," *IEEE Transactions on Power Systems*, vol. 35, 11 2020.
- [163] X. Liu, Y. Li, L. Wang, J. Tang, H. Qiu, A. Berizzi, I. Valentin, and C. Gao, "Dynamic aggregation strategy for a virtual power plant to improve flexible regulation ability," *Energy*, vol. 297, p. 131261, 2024.
- [164] B. Cui, A. Zamzam, and A. Bernstein, "Network-cognizant time-coupled aggregate flexibility of distribution systems under uncertainties," in *2021 American Control Conference (ACC)*, 2021, pp. 4178–4183.
- [165] A. Arrigo *et al.*, "Wasserstein distributionally robust chance-constrained optimization for energy and reserve dispatch: An exact and physically-bounded formulation," *Europ. Journal of Operational Research*, 2022.
- [166] X. Liu, X. Lin, H. Qiu, Y. Li, and T. Huang, "Optimal aggregation and disaggregation for coordinated operation of virtual power plant with distribution network operator," *Applied Energy*, vol. 376, p. 124142, 2024.
- [167] Y. Xu, L. Yao, T. Pu, S. Liao, F. Cheng, Y. Li, and X. Wang, "Voltage-dependent pq reserve capacity evaluation at tso-dso interface considering uncertainties of dgs and fls," *CSEE Journal of Power and Energy Systems*, 2022.
- [168] M. Zhang, X. Zhu, and N. Lu, "A data-driven probabilistic-based flexibility region estimation method for aggregated distributed energy resources," *arXiv preprint arXiv:2110.07406*, 2021.
- [169] Z. Yi, Y. Xu, W. Gu, L. Yang, and H. Sun, "Aggregate operation model for numerous small-capacity distributed energy resources considering uncertainty," *IEEE Transactions on Smart Grid*, 2021.
- [170] S. Wang, L. Du, B. Cui, and Y. Li, "Multi-factor-coupled, ahead-of-time aggregation of power flexibility under forecast uncertainty," *IEEE Transactions on Sustainable Energy*, 2024.
- [171] G. T. Heydt, "The impact of electric vehicle deployment on load management strategies," *IEEE transactions on power apparatus and systems*, no. 5, pp. 1253–1259, 1983.
- [172] S. Küfeoğlu and M. G. Pollitt, "The impact of pvs and evs on domestic electricity network charges: A case study from great britain," *Energy policy*, vol. 127, pp. 412–424, 2019.
- [173] M. Wolinetz, J. Axsen, J. Peters, and C. Crawford, "Simulating the value of electric-vehicle-grid integration using a behaviourally realistic model," *Nature Energy*, vol. 3, no. 2, pp. 132–139, 2018.

- [174] R. Dufo-López, J. S. Artal-Sevil, J. L. Bernal-Agustín, and J. A. Domínguez-Navarro, "Can electric vehicles reduce electricity bill?" in *2014 IEEE International Electric Vehicle Conference (IEVC)*. IEEE, 2014, pp. 1–4.
- [175] B. Caulfield, D. Furszyfer, A. Stefaniec, and A. Foley, "Measuring the equity impacts of government subsidies for electric vehicles," *Energy*, vol. 248, p. 123588, 2022.
- [176] K. Canepa, S. Hardman, and G. Tal, "An early look at plug-in electric vehicle adoption in disadvantaged communities in california," *Transport Policy*, vol. 78, pp. 19–30, 2019.
- [177] G. Bauer, C.-W. Hsu, and N. Lutsey, "When might lower-income drivers benefit from electric vehicles? quantifying the economic equity implications of electric vehicle adoption," *Work. Pap.*, vol. 6, pp. 1–21, 2021.
- [178] C. Lowans, A. Foley, D. F. Del Rio, and B. K. Sovacool, "Towards more equitable energy transitions in low-income households: an integrated analysis of energy and transport poverty in northern ireland," *Energy Conversion and Management*, vol. 291, p. 117337, 2023.
- [179] M. Martiskainen, B. K. Sovacool, M. Lacey-Barnacle, D. Hopkins, K. E. Jenkins, N. Simcock, G. Mattioli, and S. Bouzarovski, "New dimensions of vulnerability to energy and transport poverty," *Joule*, vol. 5, no. 1, pp. 3–7, 2021.
- [180] National Grid ESO, "Future energy scenarios 2022," 2022. [Online]. Available: <https://www.nationalgrideso.com/future-energy/future-energy-scenarios>
- [181] O. Van Vliet, A. S. Brouwer, T. Kuramochi, M. van Den Broek, and A. Faaij, "Energy use, cost and co2 emissions of electric cars," *Journal of power sources*, vol. 196, no. 4, pp. 2298–2310, 2011.
- [182] H. W. Stanton, "Impact of plug-in hybrid vehicles on the electric grid," *Oak Ridge National Laboratory, Tennessee*, 2006.
- [183] Z. Darabi and M. Ferdowsi, "Impact of plug-in hybrid electric vehicles on electricity demand profile," *Smart Power Grids 2011*, pp. 319–349, 2012.
- [184] R. A. Verzijlbergh, Z. Lukszo, J. G. Slootweg, and M. D. Ilic, "The impact of controlled electric vehicle charging on residential low voltage networks," in *2011 International Conference on Networking, Sensing and Control*. IEEE, 2011, pp. 14–19.
- [185] S. Syed Mustafa, M. Mohd Yusoff, N. Baharin, and T. Tuan Abdullah, "Modeling the impact of phev penetration on the residential electrical networks," *Applied Mechanics and Materials*, vol. 793, pp. 154–158, 2015.
- [186] K. Schneider, C. Gerkensmeyer, M. Kintner-Meyer, and R. Fletcher, "Impact assessment of plug-in hybrid vehicles on pacific northwest distribution systems," in *2008 IEEE power and energy society general meeting-conversion and delivery of electrical energy in the 21st century*. IEEE, 2008, pp. 1–6.
- [187] P. Papadopoulos, S. Skarvelis-Kazakos, I. Grau, L. M. Cipcigan, and N. Jenkins, "Electric vehicles' impact on british distribution networks," *IET Electrical Systems in Transportation*, vol. 2, no. 3, pp. 91–102, 2012.

- [188] M. Neaimeh, R. Wardle, A. M. Jenkins, J. Yi, G. Hill, P. F. Lyons, Y. Hübner, P. T. Blythe, and P. C. Taylor, "A probabilistic approach to combining smart meter and electric vehicle charging data to investigate distribution network impacts," *Applied Energy*, vol. 157, pp. 688–698, 2015.
- [189] R. A. Verzijlbergh, M. O. Grond, Z. Lukszo, J. G. Slootweg, and M. D. Ilic, "Network impacts and cost savings of controlled ev charging," *IEEE transactions on Smart Grid*, vol. 3, no. 3, pp. 1203–1212, 2012.
- [190] G. Allan, I. Eromenko, M. Gilmartin, I. Kockar, and P. McGregor, "The economics of distributed energy generation: A literature review," *Renewable and Sustainable Energy Reviews*, vol. 42, pp. 543–556, 2015.
- [191] M. Casisi, A. De Nardi, P. Pinamonti, and M. Reini, "Effect of different economic support policies on the optimal synthesis and operation of a distributed energy supply system with renewable energy sources for an industrial area," *Energy Conversion and Management*, vol. 95, pp. 131–139, 2015.
- [192] J. Liu and C. Zhong, "An economic evaluation of the coordination between electric vehicle storage and distributed renewable energy," *Energy*, vol. 186, p. 115821, 2019.
- [193] M. A. Abdullah, K. M. Muttaqi, and A. P. Agalgaonkar, "Sustainable energy system design with distributed renewable resources considering economic, environmental and uncertainty aspects," *Renewable Energy*, vol. 78, pp. 165–172, 2015.
- [194] J. Wang, "Renewable energy, inequality and environmental degradation," *Journal of Environmental Management*, vol. 356, p. 120563, 2024.
- [195] G. Konzen, R. Best, and N. J. de Castro, "The energy injustice of household solar energy: A systematic review of distributional disparities in residential rooftop solar adoption," *Energy Research & Social Science*, vol. 111, p. 103473, 2024.
- [196] C. Ramos, M. Alvargonzález, and B. Moreno, "Study of energy poverty in the european union: the effect of distributed generation," *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 17, no. 1, p. 2018525, 2022.
- [197] M. Pagliaro and F. Meneguzzo, "Distributed generation from renewable energy sources: ending energy poverty across the world," *Energy Technology*, vol. 8, no. 7, p. 2000126, 2020.
- [198] A. Bhide and C. R. Monroy, "Energy poverty: A special focus on energy poverty in india and renewable energy technologies," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 2, pp. 1057–1066, 2011.
- [199] S. Wilkinson, M. J. Maticka, Y. Liu, and M. John, "The duck curve in a drying pond: The impact of rooftop pv on the western australian electricity market transition," *Utilities Policy*, vol. 71, p. 101232, 2021.
- [200] K. B. Debnath, D. P. Jenkins, S. Patidar, and A. D. Peacock, "Remote work might unlock solar pv's potential of cracking the 'duck curve'," *Applied Energy*, vol. 367, p. 123378, 2024.

- [201] N. Etherden and M. H. Bollen, "Overload and overvoltage in low-voltage and medium-voltage networks due to renewable energy—some illustrative case studies," *Electric Power Systems Research*, vol. 114, pp. 39–48, 2014.
- [202] J. Petinrin and M. Shaabanb, "Impact of renewable generation on voltage control in distribution systems," *Renewable and Sustainable Energy Reviews*, vol. 65, pp. 770–783, 2016.
- [203] R. Ufa, Y. Y. Malkova, V. Rudnik, M. Andreev, and V. Borisov, "A review on distributed generation impacts on electric power system," *International journal of hydrogen energy*, vol. 47, no. 47, pp. 20 347–20 361, 2022.
- [204] L. Collins and J. Ward, "Real and reactive power control of distributed pv inverters for overvoltage prevention and increased renewable generation hosting capacity," *Renewable Energy*, vol. 81, pp. 464–471, 2015.
- [205] A. Rajabi, S. Elphick, J. David, A. Pors, and D. Robinson, "Innovative approaches for assessing and enhancing the hosting capacity of pv-rich distribution networks: An australian perspective," *Renewable and Sustainable Energy Reviews*, vol. 161, p. 112365, 2022.
- [206] M. Zheng, C. J. Meinrenken, and K. S. Lackner, "Smart households: Dispatch strategies and economic analysis of distributed energy storage for residential peak shaving," *Applied Energy*, vol. 147, pp. 246–257, 2015.
- [207] B. Zakeri, G. C. Gisse, P. E. Dodds, and D. Subkhankulova, "Centralized vs. distributed energy storage—benefits for residential users," *Energy*, vol. 236, p. 121443, 2021.
- [208] J. Chen and X. Song, "Economics of energy storage technology in active distribution networks," *Journal of Modern Power Systems and Clean Energy*, vol. 3, no. 4, pp. 583–588, 2015.
- [209] M. Yebiyo, R. Mercado, A. Gillich, I. Chaer, A. Day, and A. Paurine, "Novel economic modelling of a peer-to-peer electricity market with the inclusion of distributed energy storage—the possible case of a more robust and better electricity grid," *The Electricity Journal*, vol. 33, no. 2, p. 106709, 2020.
- [210] D. P. Brown, "Socioeconomic and demographic disparities in residential battery storage adoption: Evidence from california," *Energy Policy*, vol. 164, p. 112877, 2022.
- [211] W. McNamara, H. Passell, M. Montes, R. Jeffers, and I. Gyuk, "Seeking energy equity through energy storage," *The Electricity Journal*, vol. 35, no. 1, p. 107063, 2022.
- [212] G. Zubi, G. V. Fracastoro, J. M. Lujano-Rojas, K. El Bakari, and D. Andrews, "The unlocked potential of solar home systems; an effective way to overcome domestic energy poverty in developing regions," *Renewable Energy*, vol. 132, pp. 1425–1435, 2019.

- [213] V. B. Costa, R. S. Capaz, and B. D. Bonatto, "Small steps towards energy poverty mitigation: Life cycle assessment and economic feasibility analysis of a photovoltaic and battery system in a brazilian indigenous community," *Renewable and Sustainable Energy Reviews*, vol. 180, p. 113266, 2023.
- [214] R. Best and K. Sinha, "Fuel poverty policy: Go big or go home insulation," *Energy economics*, vol. 97, p. 105195, 2021.
- [215] A. Rahimi, M. Zarghami, M. Vaziri, and S. Vadhva, "A simple and effective approach for peak load shaving using battery storage systems," in *2013 North American Power Symposium (NAPS)*. IEEE, 2013, pp. 1–5.
- [216] A. Barzkar and S. M. H. Hosseini, "A novel peak load shaving algorithm via real-time battery scheduling for residential distributed energy storage systems," *International Journal of Energy Research*, vol. 42, no. 7, pp. 2400–2416, 2018.
- [217] D. Kucevic, L. Semmelmann, N. Collath, A. Jossen, and H. Hesse, "Peak shaving with battery energy storage systems in distribution grids: a novel approach to reduce local and global peak loads," *Electricity*, vol. 2, no. 4, pp. 573–589, 2021.
- [218] A. J. Pimm, T. T. Cockerill, and P. G. Taylor, "The potential for peak shaving on low voltage distribution networks using electricity storage," *Journal of Energy Storage*, vol. 16, pp. 231–242, 2018.
- [219] A. K. Srivastava, A. A. Kumar, and N. N. Schulz, "Impact of distributed generations with energy storage devices on the electric grid," *IEEE Systems Journal*, vol. 6, no. 1, pp. 110–117, 2012.
- [220] N. Jayasekara, M. A. S. Masoum, and P. J. Wolfs, "Optimal operation of distributed energy storage systems to improve distribution network load and generation hosting capability," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 1, pp. 250–261, 2016.
- [221] Y. Ma, M. S. S. Abad, D. Azuatalam, G. VerbiC, and A. Chapman, "Impacts of community and distributed energy storage systems on unbalanced low voltage networks," in *2017 Australasian Universities Power Engineering Conference (AUPEC)*, 2017, pp. 1–6.
- [222] O. Babacan, W. Torre, and J. Kleissl, "Siting and sizing of distributed energy storage to mitigate voltage impact by solar pv in distribution systems," *Solar Energy*, vol. 146, pp. 199–208, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0038092X17301494>
- [223] M. J. E. Alam, K. M. Muttaqi, and D. Sutanto, "Mitigation of rooftop solar pv impacts and evening peak support by managing available capacity of distributed energy storage systems," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 3874–3884, 2013.
- [224] O. B. Adewumi, G. Fotis, V. Vita, D. Nankoo, and L. Ekonomou, "The impact of distributed energy storage on distribution and transmission networks' power quality," *Applied Sciences*, vol. 12, no. 13, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/13/6466>

- [225] International Energy Agency, "The future of heat pumps," 2022. [Online]. Available: <https://www.iea.org/reports/the-future-of-heat-pumps/executive-summary>
- [226] V. Ballerini, E. Rossi di Schio, and P. Valdiserri, "How the energy price variability in Italy affects the cost of building heating: a trnsys-guided comparison between air-source heat pumps and gas boilers," *Buildings*, vol. 12, no. 11, p. 1936, 2022.
- [227] Y. Wang and W. He, "Temporospatial techno-economic analysis of heat pumps for decarbonising heating in Great Britain," *Energy and Buildings*, vol. 250, p. 111198, 2021.
- [228] E. Ruffino, B. Piga, A. Casasso, and R. Sethi, "Heat pumps, wood biomass and fossil fuel solutions in the renovation of buildings: A techno-economic analysis applied to Piedmont region (NW Italy)," *Energies*, vol. 15, no. 7, p. 2375, 2022.
- [229] V. Ballerini and E. R. di Schio, "Replacement of gas boilers with air source heat pumps: An economic analysis based on a trnsys evaluation and on the actual energy prices in Italy," *JP Journal of Heat and Mass Transfer*, vol. 30, pp. 151–160, 2022.
- [230] M. Ryland and W. He, "Heating economics evaluated against emissions: An analysis of low-carbon heating systems with spatiotemporal and dwelling variations," *Energy and Buildings*, vol. 277, p. 112561, 2022.
- [231] A. Lyden, S. Alene, P. Connor, R. Renaldi, and S. Watson, "Impact of locational pricing on the roll out of heat pumps in the UK," *Energy Policy*, vol. 187, p. 114043, 2024.
- [232] A. Katris, K. Turner, C. F. Calvillo, and L. Zhou, "The importance of heat pump cost reduction and domestic supply chain development in the presence of persisting energy price shocks," *Energy Strategy Reviews*, vol. 55, p. 101518, 2024.
- [233] J. Love, A. Z. Smith, S. Watson, E. Oikonomou, A. Summerfield, C. Gleeson, P. Bid-dulph, L. F. Chiu, J. Wingfield, C. Martin *et al.*, "The addition of heat pump electricity load profiles to GB electricity demand: Evidence from a heat pump field trial," *Applied Energy*, vol. 204, pp. 332–342, 2017.
- [234] T. Fawcett, N. Eyre, and R. Layberry, "Heat pumps and global residential heating," in *eccee 2015 Summer Study on energy efficiency*, 2015, pp. 321–331.
- [235] S. Watson, J. Crawley, K. Lomas, and R. Buswell, "Predicting future GB heat pump electricity demand," *Energy and Buildings*, vol. 286, p. 112917, 2023.
- [236] T. A. Deetjen, L. Walsh, and P. Vaishnav, "US residential heat pumps: the private economic potential and its emissions, health, and grid impacts," *Environmental Research Letters*, vol. 16, no. 8, p. 084024, 2021.
- [237] P. R. White, J. D. Rhodes, E. J. Wilson, and M. E. Webber, "Quantifying the impact of residential space heating electrification on the Texas electric grid," *Applied Energy*, vol. 298, p. 117113, 2021.
- [238] N. Damianakis, G. R. C. Mouli, P. Bauer, and Y. Yu, "Assessing the grid impact of electric vehicles, heat pumps & PV generation in Dutch LV distribution grids," *Applied Energy*, vol. 352, p. 121878, 2023.

- [239] C. McGarry, A. Anderson, I. Elders, and S. Galloway, "A scalable geospatial data-driven localization approach for modeling of low voltage distribution networks and low carbon technology impact assessment," *IEEE Access*, vol. 11, pp. 64 567–64 585, 2023.
- [240] C. McGarry, J. Dixon, J. Flower, W. Bukhsh, C. Brand, K. Bell, and S. Galloway, "Electrified heat and transport: Energy demand futures, their impacts on power networks and what it means for system flexibility," *Applied Energy*, vol. 360, p. 122836, 2024.
- [241] R. Guo, S. Meunier, and D. Saelens, "Impact of residential heat pumps and photovoltaics on low-voltage grid reinforcements under varying insulation levels," *Energy and Buildings*, p. 114436, 2024.
- [242] S. Meunier, C. Protopapadaki, R. Baetens, and D. Saelens, "Impact of residential low-carbon technologies on low-voltage grid reinforcements," *Applied Energy*, vol. 297, p. 117057, 2021.
- [243] C. Protopapadaki and D. Saelens, "Heat pump and pv impact on residential low-voltage distribution grids as a function of building and district properties," *Applied Energy*, vol. 192, pp. 268–281, 2017.
- [244] S. Elmallah, A. M. Brockway, and D. Callaway, "Can distribution grid infrastructure accommodate residential electrification and electric vehicle adoption in northern california?" *Environmental Research: Infrastructure and Sustainability*, vol. 2, no. 4, p. 045005, 2022.
- [245] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications surveys & tutorials*, vol. 18, no. 1, pp. 732–794, 2015.
- [246] A. Radovanovic, B. Chen, S. Talukdar, B. Roy, A. Duarte, and M. Shahbazi, "Power modeling for effective datacenter planning and compute management," *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 1611–1621, 2021.
- [247] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes, "Large-scale cluster management at google with borg," in *Proceedings of the tenth european conference on computer systems*, 2015, pp. 1–17.
- [248] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, and A. Akella, "Multi-resource packing for cluster schedulers," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 455–466, 2014.
- [249] Z. Zhang, C. Li, Y. Tao, R. Yang, H. Tang, and J. Xu, "Fuxi: a fault-tolerant resource management and job scheduling system at internet scale," in *Proceedings of the VLDB Endowment*, vol. 7, no. 13. VLDB Endowment Inc., 2014, pp. 1393–1404.
- [250] W. Liu, Y. Yan, Y. Sun, H. Mao, M. Cheng, P. Wang, and Z. Ding, "Online job scheduling scheme for low-carbon data center operation: An information and energy nexus perspective," *Applied Energy*, vol. 338, p. 120918, 2023.
- [251] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proceedings of the 15th ACM workshop on hot topics in networks*, 2016, pp. 50–56.

- [252] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE communications surveys & tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.
- [253] H. Mao, M. Schwarzkopf, S. B. Venkatakrishnan, Z. Meng, and M. Alizadeh, "Learning scheduling algorithms for data processing clusters," in *Proceedings of the ACM special interest group on data communication*, 2019, pp. 270–288.
- [254] Z. Dong, N. Liu, and R. Rojas-Cessa, "Greedy scheduling of tasks with time constraints for energy-efficient cloud-computing data centers," *Journal of Cloud Computing*, vol. 4, pp. 1–14, 2015.
- [255] K. Gai, M. Qiu, and H. Zhao, "Energy-aware task assignment for mobile cyber-enabled applications in heterogeneous cloud computing," *Journal of Parallel and Distributed Computing*, vol. 111, pp. 126–135, 2018.
- [256] G. Ye, F. Gao, and J. Fang, "A mission-driven two-step virtual machine commitment for energy saving of modern data centers through ups and server coordinated optimizations," *Applied Energy*, vol. 322, p. 119467, 2022.
- [257] N. Zilberman, E. M. Schooler, U. Cummings, R. Manohar, D. Nafus, R. Soulé, and R. Taylor, "Toward carbon-aware networking," *ACM SIGENERGY Energy Informatics Review*, vol. 3, no. 3, pp. 15–20, 2023.
- [258] C. Silva, R. Vilaça, A. Pereira, and R. Bessa, "A review on the decarbonization of high-performance computing centers," *Renewable and Sustainable Energy Reviews*, vol. 189, p. 114019, 2024.
- [259] Z. Ding, L. Xie, Y. Lu, P. Wang, and S. Xia, "Emission-aware stochastic resource planning scheme for data center microgrid considering batch workload scheduling and risk management," *IEEE Transactions on Industry Applications*, vol. 54, no. 6, pp. 5599–5608, 2018.
- [260] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, 2012, pp. 175–186.
- [261] X. Deng, D. Wu, J. Shen, and J. He, "Eco-aware online power management and load scheduling for green cloud datacenters," *IEEE Systems Journal*, vol. 10, no. 1, pp. 78–87, 2014.
- [262] A. Radovanović, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, D. Xiao, M. Haridasan, P. Hung, N. Care *et al.*, "Carbon-aware computing for datacenters," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1270–1280, 2022.
- [263] S. Chen, P. Li, H. Ji, H. Yu, J. Yan, J. Wu, and C. Wang, "Operational flexibility of active distribution networks with the potential from data centers," *Appl. Energy*, vol. 293, no. March, p. 116935, 2021.

- [264] Y. Zhang, D. C. Wilson, I. C. Paschalidis, and A. K. Coskun, "HPC Data Center Participation in Demand Response: An Adaptive Policy with QoS Assurance," *IEEE transactions on sustainable computing*, vol. 7, no. 1, pp. 157–171, 2022.
- [265] Y. Cao, F. Cao, Y. Wang, J. Wang, L. Wu, and Z. Ding, "Managing data center cluster as non-wire alternative: A case in balancing market," *Applied Energy*, vol. 360, p. 122769, 2024.
- [266] Y. Cao, M. Cheng, S. Zhang, H. Mao, P. Wang, C. Li, Y. Feng, and Z. Ding, "Data-driven flexibility assessment for internet data center towards periodic batch workloads," *Applied Energy*, vol. 324, p. 119665, 2022.
- [267] P. Zhang, M. Zhou, and X. Wang, "An intelligent optimization method for optimal virtual machine allocation in cloud data centers," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 4, pp. 1725–1735, 2020.
- [268] H. Chen, Y. Zhang, M. C. Caramanis, and A. K. Coskun, "Energyqare: Qos-aware data center participation in smart grid regulation service reserve provision," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, vol. 4, no. 1, pp. 1–31, 2019.
- [269] D. Ardagna, M. Ciavotta, and M. Passacantando, "Generalized nash equilibria for the service provisioning problem in multi-cloud systems," *IEEE Transactions on Services Computing*, vol. 10, no. 3, pp. 381–395, 2015.
- [270] H. Yuan, J. Bi, M. Zhou, Q. Liu, and A. C. Ammari, "Biobjective task scheduling for distributed green data centers," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 731–742, 2020.
- [271] Alibaba Cloud, "Elastic Compute Service Service Level Agreement," Updated at: 2023-01-13, accessed: 2024-12-23. [Online]. Available: <https://www.alibabacloud.com/help/en/legal/latest/elastic-compute-service-service-level-agreement>
- [272] Google, "Compute Engine Service Level Agreement (SLA)," Last modified November 18, 2024, accessed: 2024-12-23. [Online]. Available: <https://cloud.google.com/compute/sla>
- [273] Amazon AWS, "Amazon Compute Service Level Agreement," Last Updated: May 25, 2022, accessed: 2024-12-23. [Online]. Available: https://aws.amazon.com/compute/sla/?did=sla_card&trk=sla_card
- [274] Oracle Cloud, "Oracle Cloud Infrastructure Service Level Agreement (SLA)," Last modified October 2024, accessed: 2024-12-23. [Online]. Available: <https://www.oracle.com/uk/cloud/sla/>
- [275] A. Jackson, "Meta's 2gw data centre: How the company plans to grow ai," <https://datacentremagazine.com/technology-and-ai/metas-2gw-data-centre-how-the-company-plans-to-grow-ai>, Jan. 2025, accessed: 2025-08-11.

- [276] —, “How xai is boosting its global data centre activity,” <https://datacentremagazine.com/news/how-xai-is-boosting-its-global-data-centre-activity>, Jul. 2025, accessed: 2025-08-11.
- [277] AlphaStruxure, “Before ai, after ai: Surveying the data center industry as it enters a new age of constrained energy supply,” AlphaStruxure, Tech. Rep., Apr. 2025. [Online]. Available: https://alphastruxure.com/wp-content/uploads/2025/04/AlphaStruxure_Before_AI_After_AI_Energy_Crunch_Survey.pdf
- [278] M. Gooding, “Elon musk’s xai data center adding to memphis air quality problems, campaign group,” *Data Center Dynamics*, Aug. 2024. [Online]. Available: <https://www.datacenterdynamics.com/en/news/elon-musks-xai-data-center-adding-to-memphis-air-quality-problems-campaign-group/>
- [279] W. Gao, Q. Hu, Z. Ye, P. Sun, X. Wang, Y. Luo, T. Zhang, and Y. Wen, “Deep learning workload scheduling in gpu datacenters: Taxonomy, challenges and vision,” *arXiv preprint arXiv:2205.11913*, 2022.
- [280] C. Hwang, T. Kim, S. Kim, J. Shin, and K. Park, “Elastic resource sharing for distributed deep learning,” in *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, 2021, pp. 721–739.
- [281] A. Qiao, S. K. Choe, S. J. Subramanya, W. Neiswanger, Q. Ho, H. Zhang, G. R. Ganger, and E. P. Xing, “Pollux: Co-adaptive cluster scheduling for goodput-optimized deep learning,” in *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, 2021.
- [282] Y. Wu, K. Ma, X. Yan, Z. Liu, Z. Cai, Y. Huang, J. Cheng, H. Yuan, and F. Yu, “Elastic deep learning in multi-tenant gpu clusters,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 1, pp. 144–158, 2021.
- [283] J. Gu, M. Chowdhury, Y. Shin, Kang G. and Zhu, M. Jeon, J. Qian, H. Liu, and C. Guo, “Tiresias: A GPU cluster manager for distributed deep learning,” in *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, 2019, pp. 485–500.
- [284] F. Filippini, D. Ardagna, M. Lattuada, E. Amaldi, M. Riedl, K. Materka, P. Skrzypek, M. Ciavotta, F. Magugliani, and M. Cicala, “Andreas: Artificial intelligence training scheduler for accelerated resource clusters,” in *2021 8th International Conference on Future Internet of Things and Cloud (FiCloud)*. IEEE, 2021, pp. 388–393.
- [285] F. Wang, W. Zhang, S. Lai, M. Hao, and Z. Wang, “Dynamic gpu energy optimization for machine learning training workloads,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 11, pp. 2943–2954, 2021.
- [286] H. Zheng, F. Xu, L. Chen, Z. Zhou, and F. Liu, “Cynthia: Cost-efficient cloud resource provisioning for predictable distributed deep neural network training,” in *Proceedings of the 48th International Conference on Parallel Processing*, 2019, pp. 1–11.

- [287] S. Zhang, M. Xu, W. Y. B. Lim, and D. Niyato, "Sustainable aigc workload scheduling of geo-distributed data centers: A multi-agent reinforcement learning approach," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 3500–3505.
- [288] D. Gu, Y. Zhao, P. Sun, X. Jin, and X. Liu, "Greenflow: A carbon-efficient scheduler for deep learning workloads," *IEEE Transactions on Parallel and Distributed Systems*, 2024.
- [289] Y. Li, M. Mughees, Y. Chen, and Y. R. Li, "The unseen AI disruptions for power grids: LLM-induced transients," *arXiv:2409.11416*, 2024.
- [290] H. Zhao, M. Tanneau, and P. V. Hentenryck, "On the viability of stochastic economic dispatch for real-time energy market clearing," 2024. [Online]. Available: <https://arxiv.org/abs/2308.06386>
- [291] B. Dutta, V. Adhinarayanan, and W.-c. Feng, "Gpu power prediction via ensemble machine learning for dvfs space exploration," in *Proceedings of the 15th ACM International Conference on Computing Frontiers*, 2018, pp. 240–243.
- [292] J. F. D. Guerreiro, "DVFS modeling for energy-efficient GPU computing," Ph.D. dissertation, UNIVERSIDADE DE LISBOA, Jun. 2020.
- [293] A. Krzywaniak, P. Czarnul, and J. Proficz, "Depo: A dynamic energy-performance optimizer tool for automatic power capping for energy efficient high-performance computing," *Software: Practice and Experience*, vol. 52, no. 12, pp. 2598–2634, 2022.
- [294] Z. Chen, D. Kuhn, and W. Wiesemann, "Data-driven chance constrained programs over wasserstein balls," *Operations Research*, 2022.
- [295] Y. Wang, S. Zhao, Z. Zhou, A. Botterud, Y. Xu, and R. Chen, "Risk adjustable day-ahead unit commitment with wind power based on chance constrained goal programming," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 530–541, 2016.
- [296] Z. Wu, P. Zeng, X.-P. Zhang, and Q. Zhou, "A solution to the chance-constrained two-stage stochastic program for unit commitment with wind energy integration," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4185–4196, 2016.
- [297] X. Chen, M. Sim, and P. Sun, "A robust optimization perspective on stochastic programming," *Operations research*, vol. 55, no. 6, pp. 1058–1071, 2007.
- [298] E. Tómasson, M. R. Hesamzadeh, and F. A. Wolak, "Optimal offer-bid strategy of an energy storage portfolio: A linear quasi-relaxation approach," *Applied Energy*, vol. 260, p. 114251, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261919319385>
- [299] Y. Wang, Y. Dvorkin, R. Fernández-Blanco, B. Xu, T. Qiu, and D. S. Kirschen, "Look-ahead bidding strategy for energy storage," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 3, pp. 1106–1117, 2017.
- [300] Y. Beck, I. Ljubić, and M. Schmidt, "A survey on bilevel optimization under uncertainty," *European Journal of Operational Research*, vol. 311, no. 2, pp. 401–426, 2023.

- [301] H. Heitsch, R. Henrion, T. Kleinert, and M. Schmidt, "On convex lower-level black-box constraints in bilevel optimization with an application to gas market models with chance constraints," *Journal of Global Optimization*, vol. 84, no. 3, pp. 651–685, 2022.
- [302] H. Fröh *et al.*, "Coordinated vertical provision of flexibility from distribution systems," *IEEE Transactions on Power Systems*, 2023.
- [303] J. Silva *et al.*, "Estimating the active and reactive power flexibility area at the tso-dso interface," *IEEE Transactions on Power Systems*, 2018.
- [304] M. Zhang, Y. Xu, and H. Sun, "Optimal coordinated operation for a distribution network with virtual power plants considering load shaping," *IEEE Transactions on Sustainable Energy*, 2023.
- [305] Z. Yi, Y. Xu, J. Zhou, W. Wu, and H. Sun, "Bi-level programming for optimal operation of an active distribution network with multiple virtual power plants," *IEEE Transactions on Sustainable Energy*, 2020.
- [306] K. Garifi, K. Baker, D. Christensen, and B. Touri, "Convex relaxation of grid-connected energy storage system models with complementarity constraints in DC OPF," *IEEE Transactions on Smart Grid*, 2020.
- [307] A. Thingvad, C. Ziras, J. Hu, and M. Marinelli, "Assessing the energy content of system frequency and electric vehicle charging efficiency for ancillary service provision," in *2017 52nd International Universities Power Engineering Conference (UPEC)*. IEEE, 2017, pp. 1–6.
- [308] J. Cao, D. Harrold, Z. Fan, T. Morstyn, D. Healey, and K. Li, "Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4513–4521, 2020.
- [309] A. Bernstein and E. Dall'Anese, "Linear power-flow models in multiphase distribution networks," in *ISGT-Europe*. IEEE, 2017.
- [310] A. Bernstein *et al.*, "Load flow in multiphase distribution networks: Existence, uniqueness, non-singularity and linear models," *IEEE Transactions on Power Systems*, 2018.
- [311] M. Cai, R. Yang, and Y. Zhang, "Iteration-based linearized distribution-level locational marginal price for three-phase unbalanced distribution systems," *IEEE Transactions on Smart Grid*, 2021.
- [312] S. Mhanna and P. Mancarella, "An exact sequential linear programming algorithm for the optimal power flow problem," *IEEE Transactions on Power Systems*, vol. 37, no. 1, pp. 666–679, 2021.
- [313] Z. Yang *et al.*, "A Linearized OPF Model With Reactive Power and Voltage Magnitude: A Pathway to Improve the MW-Only DC OPF," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 1734–1745, 2018.
- [314] C. Heinrich, P. Fortenbacher, A. Fuchs, and G. Andersson, "Pv-integration strategies for low voltage networks," in *ENERGYCON*, 2016.

- [315] Q. Gao, Z. Yang, W. Yin, W. Li, and J. Yu, "Internally induced branch-and-cut acceleration for unit commitment based on improvement of upper bound," *IEEE Transactions on Power Systems*, 2022.
- [316] A. Agarwal and L. Pileggi, "Large scale multi-period optimal power flow with energy storage systems using differential dynamic programming," *IEEE Transactions on Power Systems*, vol. 37, no. 3, 2022.
- [317] F. Capitanescu, "Computing cost curves of active distribution grids aggregated flexibility for tso-dso coordination," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 2381–2384, 2024.
- [318] C. Feng, Q. Chen, Y. Wang, P.-Y. Kong, H. Gao, and S. Chen, "Provision of contingency frequency services for virtual power plants with aggregated models," *IEEE Transactions on Smart Grid*, 2023.
- [319] I.-I. Avramidis, "A first "stab" at defining temporal flexibility-to-cost mapping," *IEEE Transactions on Sustainable Energy*, 2023.
- [320] K. P. Schneider *et al.*, "Analytic considerations and design basis for the IEEE distribution test feeders," *IEEE Transactions on power systems*, 2017.
- [321] T. Morstyn *et al.*, "Open: An open-source platform for developing smart local energy system applications," *Applied Energy*, vol. 275, 2020.
- [322] H. Liu, J. Qiu, and J. Zhao, "A data-driven scheduling model of virtual power plant using wasserstein distributionally robust optimization," *International Journal of Electrical Power & Energy Systems*, 2022.
- [323] S. Pfenninger and I. Staffell, "Renewables.ninja." [Online]. Available: <https://www.renewables.ninja/>
- [324] NREL, "End-Use Load Profiles for the U.S. Building Stock," 10 2021. [Online]. Available: <https://data.openei.org/submissions/4520>
- [325] M. David, M. A. Luis, and P. Lauret, "Comparison of intraday probabilistic forecasting of solar irradiance using only endogenous data," *International Journal of Forecasting*, vol. 34, no. 3, pp. 529–547, 2018.
- [326] Y. Zhang *et al.*, "Day-ahead power output forecasting for small-scale solar pv electricity generators," *IEEE Transactions on Smart Grid*, 2015.
- [327] J. G. Jetcheva, M. Majidpour, and W.-P. Chen, "Neural network model ensembles for building-level electricity load forecasts," *Energy and Buildings*, vol. 84, pp. 214–223, 2014.
- [328] J. Cheng *et al.*, "Evaluating the spatial correlations of multi-area load forecasting errors," in *PMAFS*, 2016, pp. 1–6.
- [329] M. Bazrafshan and N. Gatsis, "Comprehensive modeling of three-phase distribution systems via the bus admittance matrix," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2015–2029, 2018.

- [330] Z. Li, M. Liu, M. Xie, and J. Zhu, "Robust optimization approach with acceleration strategies to aggregate an active distribution system as a virtual power plant," *International Journal of Electrical Power & Energy Systems*, vol. 142, p. 108316, 2022.
- [331] W. Van Ackooij, "Chance constrained programming: with applications in energy management," Ph.D. dissertation, Ecole Centrale Paris, 2013.
- [332] W. Van Ackooij and W. de Oliveira, "Convexity and optimization with copulae structured probabilistic constraints," *Optimization*, vol. 65, no. 7, pp. 1349–1376, 2016.
- [333] D. Cortes, E. Fuentes, and N. Belrose, "Approximate MVN CDF," accessed: 2024-02-08. [Online]. Available: <https://approxcdf.readthedocs.io/en/latest/>
- [334] Scottish Government, Communities Analysis Division, "Scottish house condition survey, 2012-2019." UK Data Service, 2022.
- [335] Welsh Government, "Fuel poverty estimates for wales," 2022. [Online]. Available: <https://www.gov.wales/fuel-poverty-estimates-wales>
- [336] Department for Energy Security and Net Zero, "Fuel poverty methodology handbook 2025: Low income low energy efficiency (lilee)," 2025. [Online]. Available: https://assets.publishing.service.gov.uk/media/67e3d47bdcd2d93561195be6/Methodology_Handbook_2025.pdf
- [337] Department for Business, Energy and Industrial Strategy, "Quarterly energy prices," 2023. [Online]. Available: <https://www.gov.uk/government/statistical-data-sets/annual-domestic-energy-price-statistics>
- [338] —, "Consumer prices index UK," 2023. [Online]. Available: <https://www.gov.uk/government/statistical-data-sets/monthly-domestic-energy-price-statistics>
- [339] Nottingham Energy Partnership, "Energy cost comparison," 2023. [Online]. Available: <https://nottenergy.com/resources/energy-cost-comparison/>
- [340] Office for National Statistics, "Census 2021," 2021. [Online]. Available: <https://www.ons.gov.uk/census>
- [341] M. Robinson, "Boiler service cost 2023: UK average annual boiler service costs," 2023, accessed: 16-07-2023. [Online]. Available: <https://www.boilercentral.com/boiler-costs/boiler-service-price-comparison/>
- [342] Calculator.net, "Amortization calculator," 2023, accessed: 16-07-2023. [Online]. Available: <https://www.calculator.net/amortization-calculator.html?loanamount=5%2C000&loanterm=15&loantermmonth=0&cinterestrate=3&cstartmonth=6&cstartyear=2023&cexma=0&cexmsm=6&cexmsy=2023&cexya=0&cexysm=6&cexysy=2023&cexoa=0&cexosm=6&cexosy=2023&caot=0&xa1=0&xm1=6&xy1=2023&xa2=0&xm2=6&xy2=2023&xa3=0&xm3=6&xy3=2023&xa4=0&xm4=6&xy4=2023&xa5=0&xm5=6&xy5=2023&xa6=0&xm6=6&xy6=2023&xa7=0&xm7=6&xy7=2023&xa8=0&xm8=6&xy8=2023&xa9=0&xm9=6&xy9=2023&xa10=0&xm10=6&xy10=2023&printit=0>

- [343] WIKIPEDIA, "Amortizing loan - Equated monthly installment," 2023, accessed: 02-11-2023. [Online]. Available: https://en.wikipedia.org/wiki/Amortizing_loan
- [344] O. Schmidt, S. Melchior, A. Hawkes, and I. Staffell, "Projecting the future levelized cost of electricity storage technologies," *Joule*, vol. 3, no. 1, pp. 81–100, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S254243511830583X>
- [345] Department for Energy Security and Net Zero, "Greenhouse gas reporting: conversion factors 2020," 2020. [Online]. Available: <https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2020>
- [346] National Records of Scotland, "Scotland's census 2011 results," 2023. [Online]. Available: <https://www.scotlandscensus.gov.uk/search-the-census#/>
- [347] A. Canet, M. Qadrdan, N. Jenkins, and J. Wu, "Spatial and temporal data to study residential heat decarbonisation pathways in england and wales," *Scientific Data*, vol. 9, no. 1, p. 246, 2022.
- [348] Element Energy and EA Technology, "Lv network capacity study: Phase 1 report – qualitative assessment of non-conventional solutions," Department for Energy Security and Net Zero and Department for Business, Energy & Industrial Strategy, Tech. Rep., August 2022. [Online]. Available: <https://www.gov.uk/government/publications/low-voltage-network-capacity-study>
- [349] G. Wilson and N. Godfrey, "Electrical half hourly raw and cleaned datasets for Great Britain from 2009-11-05," Jun. 2020, Grant funding as part of Research Councils (UK) EP/L024756/1 - UK Energy Research Centre research programme Phase 3 Grant funding as part of Research Councils (UK) EP/V012053/1 - The Active Building Centre Research Programme (ABC RP). [Online]. Available: <https://doi.org/10.5281/zenodo.3884859>
- [350] Northern Power Grid, "Network data," 2023. [Online]. Available: <https://www.northernpowergrid.com/demand-availability-map>
- [351] RIIO-ED2 Team, "Riio-ed2 draft determinations – core methodology document," 2022. [Online]. Available: <https://www.ofgem.gov.uk/publications/riio-ed2-draft-determinations>
- [352] Department for Energy Security and Net Zero, "Desnz public attitudes tracker: Heat and energy in the home, summer 2023, UK," 2023. [Online]. Available: <https://www.gov.uk/government/statistics/desnz-public-attitudes-tracker-summer-2023>
- [353] Department for Business, Energy and Industrial Strategy, "RHI monthly installations," 2018, accessed: 02-11-2023. [Online]. Available: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fassets.publishing.service.gov.uk%2Fgovernment%2Fuploads%2Fsystem%2Fuploads%2Fattachment_data%2Ffile%2F749544%2FRHI_monthly_official_statistics_tables_Sept_2018_publication.xlsx&wdOrigin=BROWSELINK

- [354] —, “RHI Deployment Data,” 2022, accessed: 02-11-2023. [Online]. Available: <https://www.gov.uk/government/statistics/rhi-monthly-deployment-data-december-2022-annual-edition>
- [355] Energy Systems Catapult, “Interim insights from heat pump performance data,” Energy Systems Catapult, Tech. Rep., 2022. [Online]. Available: <https://es.catapult.org.uk/news/heat-pumps-shown-to-be-three-times-more-efficient-than-gas-boilers/?reportDownload=https://es.catapult.org.uk/wp-content/uploads/2023/03/EoH-Interim-Heat-Pump-Performance-Data-Analysis-Report-1.pdf>
- [356] M. Myers, E. Kourtza, D. Kane, and S. Harkin, “The cost of installing heating measures in domestic properties,” 2018. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/913508/cost-of-installing-heating-measures-in-domestic-properties.pdf
- [357] J. Flower, G. Hawker, and K. Bell, “Heterogeneity of UK residential heat demand and its impact on the value case for heat pumps,” *Energy Policy*, vol. 144, p. 111593, 2020.
- [358] National Association of Home Builders, “Home innovation electrification report,” 2021. [Online]. Available: <https://www.nahb.org/-/media/NAHB/nahb-community/docs/committees/construction-codes-and-standards-committee/home-innovation-electrification-report-2021.pdf>
- [359] Evergreen Energy, “How long do heat pumps last?” <https://www.evergreenenergy.co.uk/heat-pump-guides/how-long-do-heat-pumps-last/>, 2023, accessed: July 7, 2023.
- [360] N. Finance, “Loan calculator,” <https://www.nortonfinance.co.uk/loan-calculator>, 2023, accessed: July 7, 2023.
- [361] GOV.UK, “Heat pump grants increased by 50%,” 2023, accessed: 2023-11-05. [Online]. Available: <https://www.gov.uk/government/news/heat-pump-grants-increased-by-50-per-cent>
- [362] —, “Apply for the boiler upgrade scheme,” 2023, accessed: 2023-07-07. [Online]. Available: <https://www.gov.uk/apply-boiler-upgrade-scheme>
- [363] O. Alabi, K. Turner, G. Figus, A. Katris, and C. Calvillo, “Can spending to upgrade electricity networks to support electric vehicles (evs) roll-outs unlock value in the wider economy?” *Energy Policy*, vol. 138, p. 111117, 2020.
- [364] J. Pearl, “Causal inference in statistics: An overview,” *Statistics Surveys*, vol. 3, no. none, pp. 96 – 146, 2009. [Online]. Available: <https://doi.org/10.1214/09-SS057>
- [365] E. Department for Business and I. Strategy, “UK rooftop solar panel behavioural research,” UK Government, Tech. Rep., November 2021. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1001896/uk-rooftop-solar-panel-behavioural-research.pdf
- [366] I. Savelli and T. Morstyn, “Better together: Harnessing social relationships in smart energy communities,” *Energy Research & Social Science*, vol. 78, p. 102125, 2021.

- [367] M. H. Abbasi, B. Abdullah, R. Castano-Rosa, M. W. Ahmad, A. Rostami, and J. Cullen, "Planning energy interventions in buildings and tackling fuel poverty: Can two birds be fed with one scone?" *Energy Research & Social Science*, vol. 93, p. 102841, 2022.
- [368] Piclo, "Piclo flex insights: Sp energy networks," 2020. [Online]. Available: <https://www.piclo.energy/blog/piclo-flex-insights-sp-energy-networks>
- [369] W. Mendieta and C. A. Cañizares, "Primary frequency control in isolated microgrids using thermostatically controllable loads," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 93–105, 2021.
- [370] Project LEO Partners and TRANSITION Team, "Project leo and transition market trials report (period 2)," Project LEO Reports, 2022. [Online]. Available: <https://project-leo.co.uk/reports/project-leo-and-transition-market-trials-report-period-2/>
- [371] Project LEO, "Smart flex heat pump trial case study," Project LEO Case studies, 2023. [Online]. Available: <https://project-leo.co.uk/case-studies/smart-flex-heat-pump-trial-case-study/>
- [372] K. MacLean, R. Sansom, T. Watson, and R. Gross, "Managing heat system decarbonisation. comparing the impacts and costs of transitions in heat infrastructure. final report." Centre for Energy Policy and Technology, Imperial College London, Tech. Rep., 2016.
- [373] M. Fulwood, "A new global gas order? (part 1): The outlook to 2030 after the energy crisis," The Oxford Institute for Energy Studies, Tech. Rep., 2023. [Online]. Available: <https://www.oxfordenergy.org/publications/a-new-global-gas-order-part-1-the-outlook-to-2030-after-the-energy-crisis/>
- [374] D. Scholten, M. Bazilian, I. Overland, and K. Westphal, "The geopolitics of renewables: New board, new game," *Energy Policy*, vol. 138, p. 111059, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0301421519306469>
- [375] H. Government, "Heat pump investment roadmap - leading the way to net zero," 2023, accessed: 30-10-2023. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1166439/heat-pumps-investment-roadmap.pdf
- [376] GOV.UK, "Thousands of heat pump installers to be trained for fraction of the price," 2023, accessed: 30-10-2023. [Online]. Available: <https://www.gov.uk/government/news/thousands-of-heat-pump-installers-to-be-trained-for-fraction-of-the-price>
- [377] F. Burns, S. Strachan, and T. Clark, "Heat pump use in Scotland: an evidence review," ClimateXChange, Tech. Rep., 2021. [Online]. Available: <https://era.ed.ac.uk/handle/1842/37698>
- [378] S. J. Darby, "Smart electric storage heating and potential for residential demand response," *Energy Efficiency*, vol. 11, no. 1, pp. 67–77, 2018.
- [379] —, "Demand response and smart technology in theory and practice: Customer experiences and system actors," *Energy Policy*, vol. 143, p. 111573, 2020.

- [380] P. Wagener, "Comfort & climate box – towards better integration of heat pumps and storage. final report." International Energy Agency, Tech. Rep., 2022. [Online]. Available: <https://heatpumpingtechnologies.org/publications/final-report-part-3-annex-55-comfort-climate-box-towards-better-integration-of-heat-pumps-and-storage/>
- [381] T. Fawcett, "Identifying the limits of the rebound effect," *BIEE Conference 2018: Consumers at the Heart of the Energy System?*, 2020. [Online]. Available: <https://www.biee.org/wp-content/uploads/Fawcett-Identifying-the-limits-of-the-rebound-effect.pdf>
- [382] The UK Government, "Levelling up the united kingdom," 2023, accessed: 2023-11-05. [Online]. Available: <https://www.gov.uk/government/publications/levelling-up-the-united-kingdom>
- [383] "An Interactive API for Scaling the Estimated DaCe Flexibility and Cost," <https://colab.research.google.com/drive/1fkMINZtjUmjEtjBaS7jgH59GsJYsT0ov?usp=sharing>, accessed: July 1, 2024.
- [384] "Ancillary Services," PJM Interconnection, accessed: January 28, 2024. [Online]. Available: <https://www.pjm.com/markets-and-operations/ancillary-services.aspx>
- [385] A. Verma *et al.*, "Large-scale cluster management at google with borg," in *Proceedings of the 10th european conference on computer systems*, 2015, pp. 1–17.
- [386] "Run:AI Platform," accessed: 2023-10-22. [Online]. Available: <https://pages.run.ai/hubfs/PDFs/RunAI-Platform-vs-Kubernetes.pdf>
- [387] A. Krzywaniak, P. Czarnul, and J. Proficz, "Dynamic gpu power capping with online performance tracing for energy efficient gpu computing using depo tool," *Future Generation Computer Systems*, vol. 145, pp. 396–414, 2023.
- [388] P. Czarnul, J. Proficz, and A. Krzywaniak, "Energy-aware high-performance computing: survey of state-of-the-art tools, techniques, and environments," *Scientific Programming*, vol. 2019, no. 1, p. 8348791, 2019.
- [389] Run:ai, "Allocation of gpu fractions," <https://docs.run.ai/v2.13/Researcher/scheduling/fractions/>, 2024, accessed on 26 Jan 2024.
- [390] NVIDIA, "NVIDIA Multi-Instance GPU: Seven independent instances in a single gpu," <https://www.nvidia.com/en-gb/technologies/multi-instance-gpu/>, 2024.
- [391] Q. Hu, Z. Ye, Z. Wang, G. Wang, M. Zhang, Q. Chen, P. Sun, D. Lin, X. Wang, Y. Luo *et al.*, "Characterization of large language model development in the datacenter," in *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, 2024, pp. 709–729.
- [392] Z. Zhao, E. Rrapaj, S. Bhalachandra, B. Austin, H. A. Nam, and N. Wright, "Power analysis of nersc production workloads," in *Proceedings of the SC'23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, 2023, pp. 1279–1287.
- [393] PassMark Software, "CPU Benchmarks," https://www.cpubenchmark.net/cpu_list.php, [Accessed 29-08-2024].

- [394] —, “I don’t understand the results. What do all these numbers mean?” https://www.passmark.com/support/performance_test_faq/understanding-results.php, [Accessed 29-08-2024].
- [395] “Deep Learning GPU Benchmarks,” <https://lambdalabs.com/gpu-benchmarks>, accessed: June 6, 2024.
- [396] “Introducing Run:ai’s CPU Scheduling: Improved Productivity and Utilization for CPU-only Clusters,” Run:AI, <https://www.run.ai/blog/introducing-run-ai-cpu-scheduling-improved-productivity-and-utilization-for-cpu-only-clusters#:~:text=Enhanced%20Cluster%20Utilization%3A%20Run%3Aai’s,memory%20unit%20is%20leveraged%20effectively.>, accessed: March 4, 2024.
- [397] “GitHub InternLM/AcmeTrace,” <https://github.com/InternLM/AcmeTrace>, [Accessed 11-06-2024].
- [398] M. Jeon, S. Venkataraman, A. Phanishayee, J. Qian, W. Xiao, and F. Yang, “Analysis of large-scale multi-tenant gpu clusters for dnn training workloads,” in *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, 2019, pp. 947–960.
- [399] “GitHub - msr-fiddle/philly-traces,” <https://github.com/msr-fiddle/philly-traces>, [Accessed 11-06-2024].
- [400] Q. Weng, W. Xiao, Y. Yu, W. Wang, C. Wang, J. He, Y. Li, L. Zhang, W. Lin, and Y. Ding, “Mlaas in the wild: Workload analysis and scheduling in large-scale heterogeneous gpu clusters,” in *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, 2022, pp. 945–960.
- [401] Alibaba, “Alibaba Cluster Trace Program,” <https://github.com/alibaba/clusterdata>, [Accessed 11-06-2024].
- [402] “GitHub - S-Lab-System-Group/HeliosData: Helios Traces from SenseTime,” <https://github.com/S-Lab-System-Group/HeliosData>, [Accessed 11-06-2024].
- [403] “SMC 2021 Data Challenge: Analyzing Resource Utilization and User Behavior on Titan Supercomputer,” <https://doi.ccs.ornl.gov/ui/doi/334>, [Accessed 11-06-2024].
- [404] F. Wang, S. Oral, S. Sen, and N. Imam, “Learning from five-year resource-utilization data of titan system,” in *2019 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2019, pp. 1–6.
- [405] Wikipedia, “Mira (supercomputer),” [https://en.wikipedia.org/wiki/Mira_\(supercomputer\)](https://en.wikipedia.org/wiki/Mira_(supercomputer)), [Accessed 11-06-2024].
- [406] “ALCF Public Data,” <https://reports.alcf.anl.gov/data/>, [Accessed 11-06-2024].
- [407] “Theta/ThetaGPU | Argonne Leadership Computing Facility,” <https://www.alcf.anl.gov/alcf-resources/theta>, [Accessed 11-06-2024].
- [408] “Theta Supercomputer Set to Retire: A Look Back at Its Impact on Science at Argonne and Beyond,” <https://www.hpcwire.com/off-the-wire/theta-supercomputer-set-to-retire-a-look-back-at-its-impact-on-science-at-argonne-and-beyond/>, [Accessed 11-06-2024].

- [409] CIEMAT, "CIEMAT - Resumen Anual 2022," https://www.ciemat.es/informesanuales/resumen_anual2022.pdf, [Accessed 11-06-2024].
- [410] D. G. Feitelson and D. Tsafir, "Parallel Workloads Archive: Logs," <https://www.cs.huji.ac.il/labs/parallel/workload/logs.html>, [Accessed 11-06-2024].
- [411] "User Statistics," <https://www.hpc.cineca.it/our-activities/services/service-statistics/user-statistics/>, [Accessed 11-06-2024].
- [412] "The Aspen Institute Italia 2024 Award to Research Conducted With Super Computers | Leonardo Pre-exascale Supercomputer," <https://leonardo-supercomputer.cineca.eu/the-aspen-institute-italia-2024-award-to-research-conducted-with-super-computers/>, [Accessed 11-06-2024].
- [413] S. Erotokritou, "PRACE 22nd Call for Proposals Supports New Innovative Research and Introduces a New Peer-Review Platform," <https://prace-ri.eu/prace-22nd-call-for-proposals-supports-new-innovative-research-and-introduces-a-new-peer-review-platform/>, [Accessed 11-06-2024].
- [414] "CINECA's Marconi100 Supercomputer Accelerates Drug Discovery at University of Nottingham," <https://www.hpcwire.com/off-the-wire/cinecas-marconi100-supercomputer-accelerates-drug-discovery-at-university-of-nottingham/>, [Accessed 11-06-2024].
- [415] F. Antici, M. Seyedkazemi Ardebili, A. Bartolini, and Z. Kiziltan, "PM100: A Job Power Consumption Dataset of a Large- Scale HPC System," Nov. 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.10127767>
- [416] "Amber - Metacentrum Documentation," <https://docs.metacentrum.cz/software/sw-list/amber/>, [Accessed 11-06-2024].
- [417] "Imbalance price modules (IP)," <https://www.netztransparenz.de/en/Balancing-Capacity/Imbalance-price/Imbalance-price-modules-IP>, [Accessed 11-06-2024].
- [418] AEMO, "Market Data Nemweb," <https://visualisations.aemo.com.au/aemo/nemweb/index.html#mms-data-model>, [Accessed 11-06-2024].
- [419] "The ESO's Demand Flexibility Service," <https://www.nationalgrideso.com/industry-information/balancing-services/demand-flexibility-service/esos-demand-flexibility-service>, accessed: 14-Jan-2024.
- [420] "On-demand GPU cloud pricing," <https://lambdalabs.com/service/gpu-cloud>, accessed: September 2, 2024.
- [421] Google, "Growing the internet while reducing energy consumption," <https://datacenters.google/efficiency/>, accessed: 2025-08-08.
- [422] D. C. Wilson, I. C. Paschalidis, and A. K. Coskun, "Site-wide hpc data center demand response," in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2022, pp. 1–7.
- [423] Illinois Center for a Smarter Electric Grid (ICSEG), "Ieee 24-bus system," <https://icseg.iti.illinois.edu/ieee-24-bus-system/>, 1979, accessed: 2024-10-11.

- [424] T. Hong, Glider, and P. Pinson, "Global Energy Forecasting Competition 2012 - Wind Forecasting," <https://kaggle.com/competitions/GEF2012-wind-forecasting>, 2012, accessed: 2024-10-11.
- [425] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," pp. 357–363, 2014.
- [426] H. Zhou, Y. Zhou, J. Hu, G. Yang, D. Xie, Y. Xue, and L. Nordström, "Lstm-based energy management for electric vehicle charging in commercial-building prosumers," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1205–1216, 2021.
- [427] SP Energy Networks, "Long term development statement," 2023. [Online]. Available: https://www.spenergynetworks.co.uk/pages/long_term_development_statement.aspx
- [428] Scottish & Southern Electricity Networks, "Tools and maps," 2023. [Online]. Available: <https://network-maps.ssen.co.uk/>
- [429] National Grid, "Network capacity map," 2023. [Online]. Available: <https://www.nationalgrid.co.uk/our-network/network-capacity-map-application>
- [430] UK Power Networks, "UK power networks primary substation distribution areas," 2023. [Online]. Available: https://ukpowernetworks.opendatasoft.com/explore/dataset/ukpn_primary_postcode_area/export/
- [431] Electricity North West, "Network information," 2023. [Online]. Available: <https://www.enwl.co.uk/get-connected/network-information/long-term-development-statement/>
- [432] Google Cloud, "CPU platforms," <https://cloud.google.com/compute/docs/cpu-platforms>, [Accessed 02-09-2024].
- [433] Amazon Cloud Service, "On-Demand Plans for Amazon EC2," https://aws.amazon.com/ec2/pricing/on-demand/?nc1=h_ls, [Accessed 02-09-2024].
- [434] —, "Amazon EC2 Instance types," https://aws.amazon.com/ec2/instance-types/?nc1=h_ls, [Accessed 02-09-2024].
- [435] Oracle, "Configure and estimate costs for OCI services ," <https://www.oracle.com/uk/cloud/costestimator.html>, [Accessed 02-09-2024].
- [436] Oracle Cloud Infrastructure Blog, "Bare Metal vs. Virtual Machines: Which is Best for HPC in the Cloud?" <https://blogs.oracle.com/cloud-infrastructure/post/bare-metal-vs-virtual-machines-which-is-best-for-hpc-in-the-cloud>, [Accessed 02-09-2024].