A Novel Data-driven Approach to Analyse Flexibility in Engineering Systems Design under Uncertainty

Presented by:

Cesare Caputo

Strategic Engineering Lab

Dyson School of Design Engineering

Imperial College London

Submitted on November 6th, 2023, in partial fulfilment of the

requirements for the degree of Doctor of Philosophy

Statement of Originality

I hereby declare that this thesis and the work presented as part of it is my own, unless explicitly stated, and the contributions of other have been acknowledged.

Cesare Caputo

November 6th, 2023

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

This thesis presents the development of a novel data-driven approach to analyse flexibility in complex engineering systems design under uncertainty, with a focus on Deep Reinforcement Learning. The proposed approach complements existing methodologies particularly for concept generation and design space exploration. The procedure, in contrast to others, allows for the systematic evaluation of more sources of uncertainty and flexibility enablers by exploiting the power of neural network function approximation. The validity of the methodology is incrementally tested through three primary case studies. Firstly, an infrastructure system which exhibits tensions between modular deployment and economies of scale. Results show substantial improvements in economic value compared to all benchmarks. Statistical analysis of the generated flexibility strategy confirm the development of more dynamic and creative solutions than through alternatives Secondly, the design of a waste to energy system is tackled, coordinating decision-making across different plants in the city. Results confirmed significantly improved expected value under uncertainty as compared to benchmarks. A larger decision making space can be considered through the proposed approach, further enhancing performance under uncertainty, and highlighting its value for highly dimensional problems. Accordingly, the first ever design and planning of mobile micro-grid energy systems for nomadic communities is conducted, a highly complex problem. The results show substantial improvements in economic feasibility, system sustainability, and resilience metrics. A novel computational framework for data-driven analysis and formulation of policy tools under uncertainty is also presented and applied to the mobile energy system case study in line with ambitious governmental decarbonization and energy access objectives. Finally, a distributional perspective is considered to more dynamically tailor flexibility strategies to project specific risk-tolerance levels. Several avenues for interpretability are presented, providing a roadmap for practical use. Nonetheless, further work is necessary to continue refining and validating the novel approach to maximize its future impact.

Acknowledgements

I would like to begin by expressing my heartfelt gratitude to my primary supervisor Dr. Michel-Alexandre Cardin for his unwavering support, mentorship, guidance, and patience throughout my PhD studies. It has been an honour to be your first PhD student at Imperial, and I have thoroughly enjoyed the journey together over the past 4 years. Your insightful feedback and constructive criticism have helped me grow both personally and professionally in countless ways, and I will always be grateful for the opportunities you have provided for me. You have ignited in me a passion for research and tackling complex issues of the future. Thank you in particular for your genuine interest in my success and well-being. I could not have asked for better support and communication during some difficult personal times, which I am very aware is not provided to all students. I am proud of the work we did together, and hope we get chances to collaborate in the future to continue building upon this.

I am also incredibly indebted to my secondary supervisor Dr. Antonio Del Rio Chanona for believing in me since we first spoke, and for your constant interest and support. The many open ended discussions we had on DRL and optimization were fascinating and helped me gain an understanding of the field I could have never anticipated. In times of doubt, you always helped me see issues with

more clarity and generated a new excitement in me for solving them. You also pushed me to explore the true state of the art in DRL, beyond what was required for this PhD project, which has provided me with an invaluable skillset for years to come. My only regret is that we were never able to formally teach the DRL course we built together, but I hope the slides will be useful some day!

I am grateful for the many friends I made at Imperial, even if our interactions were limited by the pandemic. I have particularly fond memories of the times in my first office space, together with Andrew Cullen, Steven Kench, Liam Yasin, Dr. Pietro Ferraro, and Andrea Lombardo, where we shared many laughs, games, and possibly happy hours. Andy, I will always be grateful for convincing me to take up paragliding – hope next time I see you we are flying high together. I am also very appreciative of the fellow PhD students in the Strategic Engineering Lab Group, Luka Malone, Nisha Saduagkan, and Kosuke Ikeya for the many interesting discussions together at Eastside. I wish you all the best for the remainder of your studies or careers, and please feel free to reach out at any point!

I want to express my deepest gratitude to my loving family, who have been a constant source of support, encouragement, inspiration, and excitement throughout my journey. Their love, patience, and understanding have allowed me to navigate the highs and the lows of these four years – and achieving this milestone would not have been possible without them. I am particularly grateful to my parents, Paolo Caputo and Paola Agostini, for providing me with every possible opportunity to succeed, and always pushing me to pursue my interest. Vi voglio bene e non vedo l'ora di festeggiare insieme! On that note, I am also very grateful for the constant support I received from both my step-parents, Praemeenah and Juan Pablo. Our discussions throughout the years will always be cherished memories, and I hope you know you have been an important part of my life and accomplishments. To my siblings, thank you for your love, support, and interest in my work the past 4 years and the rest of our time together. I am inspired by all that you have accomplished already and am very much excited to see what your next steps will be. Livio, I am especially indebted to you for dragging me back up when I was at my lowest, and for an amazing few years living together and sharing many things.

In closing, I want to finish by acknowledging the incredible support system that has been there for all the difficult moments here in London, in particular to the love of my life Maria. Words cannot express how grateful I am to have you in my life and for all the amazing memories we have shared together. I know we are only getting started and I cannot wait for what is next for us. Thanks for being so understanding throughout this whole period, you really are an amazing person, and deserve a lot of credit for getting me to the finish line. I also need to acknowledge our beloved dog Moose, who was always there to comfort me during times of doubt. Finally, I thank you for your patience in dealing with my PhD salary, I promise those times are over and we will enjoy the full perks of London soon!

In summary, I am incredibly appreciative for the amazing support I have received from all of you, which will not be forgotten, and I hope to make you proud in my future endeavours.

Cesare Caputo

November 6th, 2023

Table of Contents

Statement of Originality 2
Copyright Declaration 2
Abstract 3
Acknowledgements 4
Table of Contents 6
Nomenclature List 11
List of Figures 15
List of Tables 21
1 Introduction 22
1.1 Recognizing Uncertainty in Engineering Systems Design 22
1.2 Flexibility, Sustainability, and Resilience 24
1.3 The Role of Flexibility Moving Forward 26
1.4 The Potential for Data-Driven Approaches and Machine Learning 28
1.5 Motivation Summary 29
1.6 Research Objectives and Contributions 30

1.6.1 Summary and Research Objectives 30
1.6.2 Contributions 32
1.7 Thesis Structure and Overview 33
1.8 Publications and Collaborations 35
2 Background and Related Work 38
2.1 Flexibility in Design 38
2.1.1 Origins and Real Options 38
2.1.2 Definition and Framework 39
2.1.3 Creativity and Concept Generation 40
2.1.4 Design Space Exploration 41
2.1.5 Decision Rule Formulation 42
2.2 Machine Learning and Deep Learning 43
2.2.1 Background 43
2.2.2 Optimization and Key Features 44
2.2.3 Machine Learning for Engineering Design 45
2.3 Deep Reinforcement Learning 46
2.3.1 Background and Notation 46
2.3.2 DRL Agents 50
2.3.3 Practical Implementation 56
2.3.4 DRL for Engineering Design 57
3 Methodological Definition: Data-driven Flexibility in Design 59
3.1 Data-driven Flexibility in Design Framework 59
3.1.1 Overview 59
3.1.2 Intuitive Example Application 60
3.1.3 Uncertainty Recognition 62
3.1.4 Concept Generation 62
3.1.5 Design Space Exploration 63
3.2 Deep Reinforcement Learning Flexibility Analysis 64
3.2.1 Introduction and Background 64
3.2.2 Process Overview 66
3.3 Incremental Evaluation Procedure for Case Studies 68
3.4 Generic Engineering System Strategic Capacity Planning Models 70
3.4.1 Inflexible Baseline 70
3.4.2 Capacity Expansion with Flexibility Decision Rules 71
3.4.3 Capacity Expansion with Deep Reinforcement Learning 72
3.4.4 Uncertainty Simulation 73
3.4.5 The Value of Flexibility 74

4 Methodological Validation: Proof of Concept Infrastructure System 75
4.1 Case Study Overview 75
4.1.1 Mathematical Model and Assumptions 76
4.2 Benchmarking 78
4.2.1 Inflexible Baseline 79
4.2.2 Flexibility Decision Rules 79
4.2.3 Dynamic Programming Sample Average Approximation 80
4.3 DRL Implementation 81
4.3.1 DRL Flexibility Analysis Formulation 81
4.3.2 Algorithmic Approach 83
4.3.3 Training and Testing 86
4.4 Relative Performance Evaluation 88
4.4.1 Numerical Results 88
4.4.2 Motivations for Flexibility Based System Design 88
4.4.3 Stochastic Benchmarking 89
4.5 Understanding and Comparing DRL policies 90
4.5.1 Capacity Evolution Scenarios 90
4.5.2 Decision Rule Distribution 93
4.6 Chapter Summary and Conclusions 96
5 Methodological Expansion: Singapore Waste to Energy System 98
5.1 Case Study Overview 98
5.1.1 Model and Environment Formulation 99
5.2 Benchmarking 102
5.2.1 Centralized benchmark design 102
5.2.2 Flexibility decision rules 102
5.3 DRL Implementation 104
5.3.1 DRL Flexibility Analysis Formulation 104
5.3.2 Algorithmic Approach 106
5.3.3 Training and Testing 109
5.4 Relative Performance Analysis 111
5.4.1 Summary of Results 111
5.4.2 Out of Sample Testing 113
5.5 Computational Insights for Decision Making 115
5.5.1 Capacity Evolution Example Scenario 115
5.5.2 DRL Agent Action Probability Distribution Evolution 116
5.6 Discussion 117
5.6.1 Relative Performance 117

5.6.2 Degree of Embedded Flexibility 118
5.6.3 Implementation during Engineering System Operation 119
5.7 Chapter Summary and Conclusions 120
6 Methodological Maturation: Mobile Micro-Grid Design and Planning 122
6.1.1 Introduction 123
6.2 Case Study Overview 128
6.2.1 Mobile Micro-Grid Mathematical Model Formulation 128
6.2.2 Background and Previous Work 139
6.2.3 Herder Migration 141
6.2.4 Electricity Demand (ED) 141
6.2.5 Heating Demand (HD) 142
6.2.6 Renewable Energy Generation 143
6.3 DRL Implementation 143
6.3.1 DRL Flexibility Analysis Formulation 143
6.3.2 Algorithmic Approach 146
6.3.3 Training and Testing 149
6.4 Relative Performance Analysis 151
6.4.1 Summary 151
6.4.2 Expected Energy System Cost 152
6.4.3 Lifetime Emissions 153
6.4.4 Energy Security and Resilience 154
6.4.5 Local Context Constraints 154
6.4.6 Value of Plug and Play (VoPP) Operation 155
6.4.7 Policy Case Sensitivity Analysis 156
6.5 Interpreting and Implementing Complex DRL policies under Uncertainty 158
6.5.1 Overview of Flexible Energy System Evolution Strategy 158
6.5.2 Decision Rule Concept Generation and Analysis 159
6.5.3 EnerGer: Design Decision-support Tool 163
6.6 Discussion 164
6.6.1 Changing Economies of Scale 164
6.6.2 Decarbonisation Pathways 164
6.6.3 Plug and Play Mobile Microgrid Technical Feasibility 164
6.6.4 Limitations and Opportunities for Future Work 165
6.7 Chapter Summary and Conclusions 166
7 A Computational Framework for Data-driven Formulation of Policy-making Tools under Uncertainty 168
7.1 Case Study Overview 169
7.2 Background and Previous Work 170

7.2.1 Nomads, Climate Change, and National Energy Policy Objectives 170
7.2.2 PP Operation as an Enabler of Urbanized Migration and Nomadism 171
7.2.3 Energy Transition Progress 173
7.2.4 Policy-Case (PC) Formulation 174
7.3 DRL Implementation 175
7.3.1 DRL Flexibility Analysis Formulation 175
7.3.2 Computational Framework for Data-driven Policy-making 177
7.3.3 Algorithmic Approach 179
7.3.4 Training and Testing 181
7.4 Relative Performance Analysis 185
7.4.1 Cost, Sustainability and Resilience 185
7.4.2 Energy Access Metrics 187
7.5 Computational Insights for Policy-tool Formulation 188
7.6 Discussion 190
7.7 Chapter Summary and Conclusions 192
8 Risk-profile Integration 194
8.1 Introduction 195
8.2 Background and Formulation 197
8.3 Distributional DRL Implementation 200
8.3.1 Algorithmic Approach 200
8.3.2 Training and Testing 201
8.4 Relative Performance Analysis 203
8.4.1 Distributional DRL Agents 203
8.4.2 Benchmarking against Standard Methods 204
8.5 Computational Insights for Risk-Tolerance Integration 205
8.5.1 Comparative Decision Rule Distribution 205
8.5.2 Quantile Function Visualization 206
8.6 Chapter Summary and Conclusions 208
9 Conclusions and Research Outlooks 209
9.1 Summary of Contributions 209
9.2 Trade-offs with Alternative Methods 210
9.3 Domain Expertise Requirements ..... **Error! Bookmark not defined.**
9.4 Direction for Future Research 214
10 References 216
11 Appendices 230
11.1 Appendix I: Model assumptions for infrastructure chapter 4 case study 230
11.2 Appendix II: Hyperparameters used for DDQN agent training in chapter 4. 231

11.3 Appendix III: Model assumptions for WTE case study from chapter 5 231
11.4 Appendix IV: Hyperparameters used for TRPO agent training in chapter 5 232
11.5 Appendix V: Detailed mathematical formulation of power flow model from chapter 6 233
11.6 Appendix VI: Model assumptions and parameters for Mongolian mobile micro-grids case study from chapter 6 236
11.7 Appendix VII: Hyperparameters used for ACER agent training in chapter 6 237
11.8 Appendix VIII: Hyperparameters used for PPO agent training in chapter 7 237
11.9 Appendix IX: Hyperparameters used for IQN agent training in chapter 8 238

Nomenclature List

Acronyms and Abbreviations

A2C	Advantage Actor Critic	MDP	Markov Decision Process
ACER	Actor Critic with Prioritized Experience Replay	ML	Machine Learning
AD	Anaerobic Digestion	MLP	Multi-layer Perceptron
AI	Artificial Intelligence	MM	Minorize-Maximize Algorithm
ANN	Artificial Neural Network	MS	Mobility Scenario
BD1/2/3/4	Baseline Rigid Design 1/2/3/4	MSSP	Multi-Stage Stochastic Programming
BESS	Battery Energy Storage System	NCF	Net Cash Flow
CAPEX	Capital Investment Cost	NN	Neural Network
CDF	Cumulative Distribution Function	NPC	Net Present Cost
CNN	Convolutional Neural Network	NPV	Net Present Value
CVaR	Conditional Value at Risk	PC	Policy Case
DCF	Discounted Cash Flow	PM	Particulate Matter
DDPG	Deep Deterministic Policy Gradient	PP	Plug and Play
DGs	Distributed Generators	PPO	Proximal Policy Optimization
DNN	Deep Neural Network	PWC	Public Waste Collection

DQN	Deep Q-Network	REAP	Rural Electricity Access Project
DDQN	Double Deep Q-Network	ReLU	Rectified Linear Unit
DR	Decision Rule	RES	Renewable Energy System
DRL	Deep Reinforcement Learning	RN	Risk Neutral
DSM	Design Structure Matrix	RNN	Recurrent Neural Network
DP	Dynamic Programming	RL	Reinforcement Learning
EH	Electric Heating/Heater	ROA	Real Options Analysis
ENPC	Expected Net Present Cost	SAA	Sample Average Approximation
ENPV	Expected Net Present Value	SAC	Soft Actor Critic
ETS	Emissions Trading Scheme	SDG	Sustainable Development Goal
EoS	Economies of Scale	SGD	Stochastic Gradient Descent
ES-MDM	Engineering Systems Multiple Domain Matrix	SHS	Small Home Solar
EQN	Ensemble Quantile Networks	SP	Stochastic Programming
FD	Flexible Design	SWS	Small Wind Systems
FDR	Flexibility Decision Rules	TD	Temporal Difference
FND	Flexible Nanogrid Design	TE	Testing Environment
FW	Food Waste	TRPO	Trust Region Policy Optimization
GA	Genetic Algorithm	TUL	Total Unmet Load
GAE	Generalized Advantage Estimation	UB	Ulaanbaatar
GANs	Generative Adversarial Networks	UN	United Nations
GBM	Geometric Brownian Motion	UK	United Kingdom
GCRF	Global Challenges Research Fund	VaR	Value at Risk
GPS	Guided Policy Search	VaG	Value at Gain
HF	High Flexibility	VoF	Value of Flexibility
IQN	Implicit Quantile Networks	WTE	Waste to Energy
KL	Kullback-Leibler	LF	Low Flexibility
LFS	Load Following Strategy	LSTM	Long Short-term Memory Network

Variables and Symbols

$\mathcal{A}$	Action Space	T	Evaluation horizon (years or months)
$\mathcal{S}$	State Space	NCF	Net cash flow (USD)
$\mathcal{P}$	Transition Probability	NPC	Net present cost (USD)
$\mathcal{R}$	Reward Function	c_r	Operational costs (USD/space)
$\mathcal{C}$	Cost Function	c_l	Land leasing costs (USD/year)
$\mathcal{H}$	Expansion Cost Function	D_t	Annual parking demand (spaces)
s	State vector	p	Cost of parking (USD/year)
a	Action vector	n_0	Capacity per floor (spaces)

t	Time-step	f_0	Starting floors
π	DRL agent policy	c_0	Cost of precast construction (USD/space)
γ	MDP discount factor	f_{max}	Max floors
V	State value function	σ_{pg}	Parking demand volatility
Q	State-action value function	a_{x-y}	Parking allowance between years x and y
ξ	Uncertainty vector	d_r	Years with excess demand
θ	Installed nominal capacity	f_t	Expansion magnitude
λ	Financial discount rate	z_n	DP SAA estimator
θ	Capacity deployed / NN parameters	ε	Epsilon exploration factor [0,1]
$P(i)$	PER sampling probability	AR^t	Total annual revenues (USD)
R_D^t	Food waste collection revenue (USD)	R_E^t	Electricity generation revenue (USD)
d_i^t	Recycled food waste by sector (ton)	P_{to}	Tipping Fee (USD/ton)
E_g	Electricity Generation rate (kWh/ton)	P_e	Price of electricity (USD/ kWh)
AC^t	Annual system cost (USD)	C_{Tran}^t	Annual transport costs (USD)
C_D^t	Waste disposal cost (USD)	C_L^t	Land Leasing Cost (USD)
C_{OM}^t	Operation and maintenance cost (USD)	Cap_v	Vehicle capacity (tonnes/trip)
Cap_v	Unit fuel cost (USD/km)	D_{co_i}	Intrasectoral transport distance (km)
K	Investment constant in EoS model	D_{tr_i}	Intersectoral transport distance (km)
α	EoS factor in EoS model	Σ	residues rate for food wastes
Ω	purity rate for food wastes	ρ	Land rental fee by capacity (USD/tpd)
$P_{ul,es}$	Unmet total load for *PP microgrid* (kWh)	EV	Expected Value
$P_{dist,es}$	Total distributed electricity in *PP microgrid* (kWh)	$VoPP$	Value of Plug and Play Operation
$\theta_{PP,es}$	Total *PP microgrid* distribution capacity limits (kW)	μ	Drift for GBM
$P_{dist,es,max}$	Constraint on maximum *PP microgrid* distribution (% of ED)	σ	Volatility for GBM
P_{dist}	Active power distribution requirements per *nanogrid* (kW)	W	Weiner process stochastic deviation
L_{dist}	Distance distribution requirements per *nanogrid* (kW)	N_n	Number of community dwellings
R_{dist}	Distribution resistance requirements per *nanogrid* (Ω)	$r_{cluster}$	Nomadic community cluster radius (km)
V_{dist}	Voltage distribution requirements per *nanogrid* (V)	L_{cb}	Cabling length (km)
I_{dist}	Current distribution requirements per *nanogrid* (amp)	P_{wind}	Active Wind power (kW)
α_{cb}	Material temperature coefficient	P_{pv}	Active PV power (kW)

$P_{dl,es}$	Active power distribution losses for *PP microgrid* (kW)	P_{NEL}	Net electricity load (kW)
DL	Total distribution loss %	P_{NHL}	Net heating load (kW)
P_{grid}	Active exchanged with main grid for *PP microgrid* (kW)	ED	Electricity demand (kW)
θ_{grid}	Total *main grid* distribution capacity limits (kW)	HD	Heating demand (kW)
ρ_{grid}	Binary variable on grid availability	P_{res}	RES active power (kW)
$P_{uhl,es,ppp}$	Unmet heating load after *PP microgrid* distribution (kW)	$P_{eh,max}$	Maximum EH capacity (kW)
$P_{res,exc,pp}$	Excess electricity after *PP microgrid* distribution (kW)	η_{inv}	Inverter roundtrip efficiency (%)
$P_{coal,es}$	Power coal requirements for *PP microgrid* (kW)	$P_{ba,in,max}$	Maximum battery charging capacity (kW)
M_{coal}	Coal mass needed for *PP microgrid* (kg)	$P_{ba,out,max}$	Maximum battery discharge capacity (kW)
$\mu_{local\ coal}$	Local lignite coal energy density	E_{ba}	Battery storage level (kWh)
$\eta_{coal\ stove}$	Coal stove net efficiency (%)	$E_{ba,min}$	Minimum battery storage level (kWh)
$CO_{2_{eq.emissio}}$	Equivalent CO_2 emissions generated (tonnes)	$E_{ba,max}$	Maximum battery storage level (kWh)
EF	Weighted average emission factor	η_{batt}	Battery roundtrip efficiency (%)
C_{inv}	Total investment costs (USD)	$P_{res,PL}$	Available RES capacity post meeting nanogrid load (kW)
C_{opex}	Operational costs (USD)	$N_{deficit}$	Nanogrid units with electricity deficit
C_{grid}	Grid active power exchange costs or revenue (USD)	$N_{surplus}$	Nanogrid units with electricity surplus
C_{CO2}	Carbon emissions equivalent costs or revenue (USD)	$P_{eg,es}$	Excess *PP microgrid* electricity generation (kWh)
$C_{es,max}$	Budget constraint (USD)	L	Functional operational life (years)
CC	Communication costs (USD)	SV	Salvage value (USD)
EC	Expansion costs (USD)	θ_{sv}	Abandoned capacity (kW)
RC	Replacement costs (USD)	k	Depreciation factor
IC	Infrastructure costs (USD)	C_{coal}	Total coal costs (USD)
PEC	Power electronics costs (USD)	c_{coal}	Coal cost factor (USD/kg)
c_{cc}	Communication costs factor (USD/ kWh)	C_{om}	Total O&M costs (USD)
c_{ic}	Cabling costs factor (USD/ km)	c_{om}	O&M cost factor (USD/kW)
c_{pe}	Power electronics cost factor (USD/ kW)	C_{ul}	Total unmet load costs (USD)
K	Investment constant in EoS model	c_{ul}	Unmet load cost factor (USD/kWh)
α	EoS factor in EoS model		
ε_{pe}	Power electronics requirement ratio	C_{grid}	Total grid interaction costs (USD)

sf_{pe}	Safety factor	c_{grid}	Grid tariff (USD/kWh)
M_{batt}	Battery total weight (kg)	p_{grid}	Grid RES subsidy (USD/kWh)
M_{pe}	Power electronics total weight (kg)	M_{RES}	RES total weight (kg)
M_{cb}	Distribution cabling weight (kg)	$RE_{subsidy}$	% of CAPEX
$RE_{feed\ in\ tar}$	RE selling price (USD/kWh)	$EH_{subsidy}$	% of CAPEX
p_{CO_2}	Carbon credit (USD/ tonnes CO_2)	Z	Distributional value function
τ	IQN base distribution	β	Distortion risk measure
$\mathcal{L}_k$	Huber loss for IQN	$\delta_t^{\tau,\tau'}$	Distributional TD error
ρ_k	Quantle Huber regression loss	k	Threshold for loss calculations
α	Risk tolerance level for CVaR	N and N'	Sampling frequency for IQN

List of Figures

Figure 1.1: Flexibility, Sustainability and Resilience overlap 25
Figure 1.2: Analysis of trade-offs motivating the development of a new method in the field..... 28
Figure 2.1: Example cumulative distribution functions for conceptual baseline and flexible systems design concepts, using NPV as lifecycle performance metric[61] 39
Figure 2.2: Flexibility in Design Framework..... 40
Figure 2.3: DRL agent-environment sequential interaction framework. 47
Figure 2.4: Graphical overview of policy gradient-based DRL optimisation process. Part **i)** shows an example policy function which we would seek to optimise, highlighted by the presence of many local minima between the initialization point A and the global optimum B. The parameters for which we seek optimal values in this case are the neural network weights $\theta 0$ and $\theta 1$ shown in Part **ii)** of the figure. Finally, Part **iii)** gives an overview of the training and optimisation process of a policy neural network when randomly initialized to point A in Part **i)** and seeking to reach point B in the same..... 53
Figure 2.5: Graphical overview of gradient-free approaches in policy-based DRL 54
Figure 2.6: Overview of actor-critic method for DRL policy optimisation. 55

Figure 3.1: Data-driven Flexibility in Design Framework..... 60
Figure 3.2: Modified ES-MSM. The impact on S, F, R represents a qualitative assessment by domain experts of the perceived effect of the interaction ranked as either positive, negative, or neutral. 63
Figure 3.3: Overview of DRL decision making process on motivating case study. Note that action selection and value function estimates are simply for illustrative purposes and do not represent the actual optimised ones obtained through a complete DRL Flexibility Analysis. 65
Figure 3.4: Primary steps in DRL flexibility analysis..... 66
Figure 3.5: High-level summary of use of DRL policy in practice..... 67
Figure 3.6: Incremental procedure implemented to evaluate the various case studies in this thesis. ... 68
Figure 3.7: Comparison showing incrementally increasing complexity of case studies evaluated. Photo credit for Chapter 4 is from[187], for Chapter 5 from[188], and for Chapter 6 from[189] 69
Figure 4.1: i) Example multi-story garage; ii) Overview of tensions involved in design process. 76
Figure 4.2: Comparison of deterministic and stochastic simulated demand evolutions. The shaded blue area represents the 95% confidence interval for stochastic demand over 2000 simulated scenarios... 78
Figure 4.3: Visualization of flexibility decision rules procedures for multi-story garage case study. Decision rule variables optimised are highlighted in black for improved graphical clarity. 80
Figure 4.4: Graphical overview of DRL agent decision-making process based on Algorithm 1. 85
Figure 4.5: Learning curve for DDQN agent on garage capacity expansion case study, where the y-axis represents the ENPV obtained in one episode (20 year project life) and the x-axis the number of training decision time-steps. The solid lines show the smoothed averages using a window size of 50 episodes. The shaded area shows the standard deviations among different training runs. 87
Figure 4.6: Summary of timestep resolutions used in different stages for the proposed DRL approach on the multi-story garage case study. 88
Figure 4.7: Cumulative distribution functions showing lifetime performance of different design alternatives for the garage case study based on 2000 simulated scenarios. Vertical lines represent the ENPV for each of the methods implemented..... 89
Figure 4.8: Example capacity evolution scenario using for the Flexibility DRs (red), and DRL Flexibility analysis (green) under different stochastic realizations of demand (blue), where i) captures a relatively high demand scenario, ii) a medium demand scenario, and iii) a lower scenario. 92
Figure 4.9: Histogram for dr variable distribution of DRL agent policy over 2000 scenarios. 94
Figure 4.10: Histogram for ft variable distribution of DRL agent policy over 2000 scenarios. 95
Figure 4.11: Distribution of expansion decisions over project life across 2000 simulated scenarios for Design 2 (red) and Design 4 (green). Vertical dashed lines are their mean expansion years. 95
Figure 5.1: (a) Example stochastic scenario of FW processing demand by sector ; (b) city-wide stochastic evolutions over 2000 simulated scenarios including low, medium, and high example scenarios, where blue shaded area represents 95% confidence interval. 101

Figure 5.2: Visualization of FDR implementation to determine expansion decisions on WTE case study. DR variables are highlighted in black for improved graphical clarity. 103
Figure 5.3: DRL decision process for both the low flex (DRL-LF) and high flex (DRL-HF) implementations. Magnitude of expansion limited to 200tpd in blue boxes for low flex implementation, no other expansion levels examined. 109
Figure 5.4: Training reward evolution for first 1 million steps of DRL-LF and DRL-HF agents on WTE case study. The solid lines show the smoothed averages using a window size of 100 episodes. The shaded areas show the standard deviations among different training runs. 110
Figure 5.5: Summary of timestep resolutions from DRL agent training and testing on WTE study. . 111
Figure 5.6:Cumulative distribution functions showing lifetime performance of different design alternatives over 2000 simulations. Vertical lines with markers represent ENPV values. 112
Figure 5.7: Example distributed food waste scenario leading to demand scenarios in sectors S1-S6 (top), corresponding design capacity expansion for FDR (bottom left), DRL-LF (bottom center) and DRL-HF designs (bottom right). Legend in top plot applies to all items. 115
Figure 5.8: Action sampling probabilities for the example distributed demand scenario in this section, for DRL LF system. Actions maintaining a probability of being selected below 5 % through the project duration are not included. 116
Figure 6.1: PP operation demonstration including typical (1) ~ (3) operational scenarios (left) and performance evaluation of frequency, active power ratio, active power, and voltage, in (a)-(d), respectively, over the same time horizon (right)..... 130
Figure 6.2: Multi energy vector block diagram of herder nanogrid (left) and PP microgrid (right)... 131
Figure 6.3: Model flowchart showing complex interactions among system components and simulations. 138
Figure 6.4: Spatial boundaries for typical migration pattern (left) and example multi-ger PP configurations (right) 139
Figure 6.5: Deterministic projections **(dashed line)** and 3 randomly selected example stochastic simulated scenarios **(solid lines),** showing the average annual (over the migration season) 18-ger cluster radius in km (y-axis), over the 30-year project horizon (x-axis). The projected migration growth is assumed to be driven by the climate change effects, as shown by the deterministic projections **(dashed line),** although it is recognized the magnitude of this impact is highly uncertain both in its long-term evolution and intra-year variability, as captured by the different stochastic example scenarios **(solid lines).**..... 141
Figure 6.6: Hourly projected load (left) and seasonal long term stochastic evolutions of electricity demand (right)..... 142
Figure 6.7: Projected starting and end of project life heating demand monthly profiles for 3 scenarios. 143

Figure 6.8: a) Graphical overview of DRL system level decision making process going from microgrid state space input (left) to action selection output (right) b) Simple heuristic used for allocation of system level DRL actions to individual microgrid nodes, either the ones with either largest unmet load in previous month (top) or greatest total excess electricity generation (bottom). 146
Figure 6.9: Training episode reward evolution for ACER agent on Mongolian mobile microgrids environment. The y-axis estimated the total energy system cost per episode while the x-axis shows the number of *monthly* decision timesteps implemented for training. The solid line shows smoothed average (using a window size of 50 episodes) among 5 runs. The shaded area the standard deviation recorded across the same training runs. 150
Figure 6.10: Summary of time-step resolutions implemented for different levels of the mobile micro-grid simulations used for the proposed DRL approach in this chapter. 151
Figure 6.11: Heating supply source (left-axis) vs carbon footprint evolution (right-axis) for FD system over 2000 simulated scenarios showing the 95% confidence interval. 154
Figure 6.12: Annual distribution loss % as a function of herder PP microgrid spatial dispersion..... 156
Figure 6.13: Policy sensitivity analysis of primary design alternatives considered in this chapter based on assumptions presented in Table 11. 158
Figure 6.14: Cumulative System Capacity Expansion for the trained ACER agent averaged over 2000 episodes. Shaded areas show 95% confidence interval across scenarios. 159
Figure 6.15: Distribution of capacity adjustment timings per technology selected by ACER agent over 2000 simulated scenarios (regardless of magnitude selected). 160
Figure 6.16: Distribution of capacity adjustment magnitude per technology selected by ACER agent over 2000 simulated scenarios (regardless of timing). All plots share the same y-axis depicted on the left hand side. For all plots shown, the left bin represents the frequency of abandonment decisions, the middle bin modular capacity expansion, and the right bin expansion by the max capacity allowed by budget and mobility constraints. As the latter is rounded to the nearest number of modules, it should be noted that this may represent different magnitudes across environment episodes evaluated. 161
Figure 6.17: RES to EH System sizing ratio for 2000 simulated episodes showing the 95% confidence interval. Baseline with EH refers to BD4 from Table 8. PC1 denotes baseline government policy case (no subsidies or carbon credits). 162
Figure 6.18: Distribution of capacity expansion decisions vs shortage levels over 2000 episodes for the trained ACER agent. 162
Figure 6.19: Visualization of interface for EnerGer prototype [275]. Going from left to right, it shows the homepage, DRL model predictions/recommendations, a geospatial feature to identify nearby communities during migration, and a summary of energy usage by heating or electricity. 163
Figure 7.1: Migration pattern of herders and mobile energy access scenarios. **a**, the city urbanization process with increasing informal settlements around the suburban area (ger districts in the Ulaanbaatar, Mongolia as an example): apart from the commercial buildings, industrial factories, and

residential households that all consume energy, urbanization leads to more ger districts around the city centre without enough energy supply infrastructure. Renewable is promising in electrifying energy usage of increasing citizens, both centralized PV stations and decentralized household PV systems. The control of all these energy units can be assisted by base station enabled wireless communication, while only communication links for Gers (in red dotted lines) are shown. **b**, a similar urbanization process but in some small settlements such as small cities or villages: there are some non-ger and ger households in the residential areas, and mobile PV panels assist the energy usage in Gers. **c**, nomadic lifestyle with increasing energy access by mobile renewable PV systems. For three scenarios (shown in **a**, **b**, **c**), the pastoralists can migrate across these three scenarios, while their mobile energy supply facilities such as PV systems and small wind turbine systems have a functionality of plug-and-play operation. **d**, a simple case of plug-and-play operation with the energy unit plugging in-and-out. **e**, the state (such as frequency and voltage variables in electricity systems) evolution of plug-and-play operation shown in subfigure **d**: with some transient, the state tracks the reference. By the plug-and-play operations (**d**, **e**) of energy supply units, the energy usage pattern in global pastoralists with migration patterns (**a**, **b**, **c**) can be mitigated, leading to an efficient, green, decarbonized energy supply and a healthy living environment with reduced air pollution. 172
Figure 7.2: Uncertainty space considered for the different policy alternatives over the project 30-year horizon (a-d). PC1 is shown in (a), PC2 in (b), PC3 in (c) and PC4 in (d).Note that for each PC the shaded area represents the 95% confidence interval (assuming a normal distribution), the solid line the mean expected value over time, and the other lines 3 randomly selected example scenarios for each PC to highlight variability across GBM uncertainty realizations. 177
Figure 7.3: **(a)** Data-driven framework for enabling the evaluation and refinement of different policy alternatives to accelerate the transition to SDG7 under a high degree of uncertainty on energy system operation. **(b)** DRL agent iterative training process for energy system planning optimisation going from entire state space inputs (left) to selected action in a particular time step (right). The multi-energy vector block diagram shown on the bottom half of the figure represents, going from left to right, an example individual nano-grid unit, the PP nomadic community microgrid. Engagement with the centralised power grid is available only during the scenarios captured by Figure 7.1a and Figure 7.1b. The uncertainty in PCs introduced into the state vector is marked by the red variables (left). Note that, in fact, **(b)** is overall similar to Figure 6.8a, although the former highlights the increased state space considered in this Chapter through the inclusion of policy uncertainty. Additionally, while the heuristic for action node selection from Figure 6.8b is not shown here to enhance visual clarity, it is implemented exactly in the same manner during energy system operation. 178
Figure 7.4: Training episode reward evolution for PPO agent on Mongolian energy system policy tool formulation environment. The y-axis estimated the total energy system cost per episode while the x-axis shows the number of *monthly* decision timesteps implemented for training. The solid line shows

smoothed average (using a window size of 100 episodes) among 5 runs. The shaded area the standard deviation recorded across the same training runs. 183
Figure 7.5:Summary of time-step resolutions implemented at different levels for DRL agent training and testing on the mobile microgrid model considered in this chapter..... 185
Figure 7.6: Multi-objective **(a-c)** evaluation of mobile and flexible energy system solutions deployed under different PCs (using their respective testing environments) capturing their expected performance under uncertainty outside of energy access metrics. The distribution of economic feasibility is captured by the boxplots in (**a**), system sustainability in (**b**), and energy security in (**c**). None that for **(a-c)** pictured above, the boxes represent the interquartile ranges under each PC for that metric, while the whiskers show the Value at Gain (**VaG**) and Value at Risk (**VaR**), or the values at the **95%** and **5%** percentiles of the performance distribution, respectively, over the testing scenarios analysed under uncertainty. The **mean** or expected value for each metric is shown by the **dotted black line in each box**, adopting the same colours as Figure 7.7 for consistency. 187
Figure 7.7: Performance of nomadic mobile energy supply system with the implementation of different PCs under uncertainty on energy access metrics SDG7.1.1 (**a**), 7.1.2 (**b**.) 7.2 (**c**) and 7.b.1 (**d**). The shaded area in between PC0 (orange line) and the best performing policy tool (stated) for each energy access metric (a)-(d) represents the cumulative percentage improvement generated by that PC..... 188
Figure 8.1: Visualization of difference between value function in classical RL and value distribution in distributional RL 197
Figure 8.2: Comparison of distribution focus for standard vs risk sensitive DRL, with shaded region representing focus and return performance metrics used in risk sensitive DRL..... 199
Figure 8.3: Learning curve for IQN agent on garage case study using three different levels of risk sensitivity. 202
Figure 8.4: Summary of timestep resolutions across training and testing process used for risk-profile integration on garage case study. 203
Figure 8.5: Cumulative distribution function showing expected performance of IQN agent with different risk-tolerance levels on the garage case study across 2000 simulated scenarios. Vertical dashed lines represent the ENPV for each agent. 204
Figure 8.6: Example quantile functions at year 1 (a), year 20 (b), year 5 (c) and year 15 (d) showing estimated returns (y-axis) for each action in a single stochastic scenario..... 207

List of Tables

Table 1: Overview of research motivation driving this thesis, with the works cited further explored and discussed in Chapter 2. 30
Table 2: Decision rules and subset of values considered in the design vector for the multi-story parking garage case study. 79
Table 3: Economic performance and VoF from the solutions in a deterministic demand scenario only, with all values shown in USD Million units. 88
Table 4: Multi-criteria design decision making assessment for the garage case study with stochastic demand over 2000 simulated scenarios. 90
Table 5: Decision rules variables and their ranges considered for the WTE case study. 103
Table 6: Design decision making table for WTE system (all values in Million S$) 112
Table 7:Out of sample testing of performance of design alternatives considered in this study. 114
Table 8: Summary of nomadic ger energy system design alternatives evaluated in this study. 151
Table 9: Design decision making table for alternatives evaluated in this study over 2000 out of sample scenarios 152

Table 10: Estimated VoPP and externalities on energy security and sustainability for different system design alternatives..... 155
Table 11: Summary of policy cases investigated for sensitivity analysis of design alternatives considered. 157
Table 12: Summary of Policy Cases (PCs) analysed, and the ranges considered for DRL agent training. Note that these were refined from the ones in 6.4.7 in line with SDG7 objectives. 174
Table 13: Summary of testing environments used for nomadic policy-tool formulation objectives in this chapter. 184
Table 14: Assumptions and parameters used in WTE model. All values are taken from the study by Cardin and Hu for benchmarking purposes. 231
Table 15: Parameters and assumptions used to build model in chapter 6. 236

Chapter 1

1 Introduction

1.1 Recognizing Uncertainty in Engineering Systems Design

Engineering design can be defined as "establishing and defining solutions or pertinent structures for problems not solved before, as well as new solutions to problems which have previously been solved in a different way"[1]. It is an essential phase in an engineering project, as decision made during design can lead to major impacts on a products overall cost, with more limited opportunities for change after the initial stages[2]. Recent years have been marked by evolving challenges in the design and planning of engineering systems which provide critical services to society such as transportation, power, healthcare, and telecommunications, among others. These complexities naturally arise from the changing operational parameters required by these systems throughout their long lifetime. In turn, they can impact performance across several important metrics through systemic dependencies between markets, policy, regulators, and technology[3]. More formally, these impacts may be driven by

both the probability of making mistakes in the assumptions during the design phase (i.e. material properties, component dependencies) or from external events which impact final system performance (i.e. market dynamics, political developments, climate change).

Climate change, pandemics, evolving economies, terrorism, and innovation are just a few sources of uncertainty an engineering system and design framework of the future must be prepared to deal with - typically classified as either epistemic or aleatoric [4]. Epistemic uncertainty is primarily concerned with inputs to the model and subsequent propagation of errors through system boundaries, thus can be minimized through appropriate feasibility studies and increasing data availability on a specific system. Aleatoric uncertainty, on the other hand, is *exogenous* to the engineering system and as a result may be irreducible, given its realization is outside the control of system designers and operators.

Uncertainty can impact the performance of an engineering system negatively or positively, depending on the evolution of the most essential variables [5]. Advances in probability theory, for instance, have been found to yield improved predictive accuracy for these variables and events – particularly through the development of data-driven methods. Uncertainty in the context of an engineering design problem may in fact be defined as making up the difference between the knowledge available to decision makers and that necessary to optimally plan the system[6]. As such, it may also impact the degree and simplicity of adapting a future system based on changing requirements or operational parameters.

Complex engineering systems design must inherently address uncertainty, but an important issue motivating this study is that current approaches are not leading to readily flexible and adaptable systems- increasingly important in today's world. System designs are often optimized using deterministic assumptions for market, environmental conditions, regulations, and technologies, making them rapidly sub-optimal if operational conditions change. Industry often invests vast efforts to take advantage of economies of scale, thereby committing large resources upfront to reduce unit costs, while such resources may not be fully utilized in the future depending on market realizations. Such strategies often lead to overly rigid engineering systems, which do not make best use of limited resources, and may be unable to quickly adapt and reconfigure following important disruptions.

These issues have given rise to a number of underperforming engineering systems – even important failures – in the face of uncertainty, despite well-functioning technologies. Consider for instance the development of the Iridium satellite cell phone system[7] developed in the 1990s by Motorola ® to provide wireless cell-phone service anywhere on the planet. Decision makers failed to recognize market uncertainty in its deployment, assuming everyone would want to switch to the Iridium product. Based on a deterministic analysis, the full capacity for the system was deployed right away at a very high total cost. Ultimately, the anticipated demand did not even close to materialize as land-based cellphones grew increasingly popular, leading the system to be sold in bankruptcy for a fraction

of the initial investment costs. Many other similar examples exist, such as IUT Global in Singapore [8], Ghost Cities [9] in China, or the high-speed railway (i.e. HS2) in the UK [10] – among others.

1.2 Flexibility, Sustainability, and Resilience

Flexibility has recently emerged as a potentially unifying and powerful paradigm to address the impact of uncertainty on the design process, allowing a system to capture more upside opportunities while reducing downside exposure[11]. The definition of flexibility used in this thesis, inspired from real options theory, is the "right, but not the obligation, to change a system in the face of uncertainty"[12,13]. This definition, therefore, extends past the technical domain to consider managerial decision-making as an enabler of valuable flexibility during operations. The higher the associated level of uncertainty, generally the more enhanced the value of embedding flexibility into the system is in terms of expected performance[14]. Flexibility, therefore, changes the objective from optimizing for a fixed "design point" operation to solutions that maintain strong performance across the entire distribution of outcomes.

The explicit consideration of uncertainty provided through flexibility may also have important implications for the integration of sustainability and resilience objectives. Sustainability is most commonly defined as a "development that meets the needs of the present without compromising the ability of future generations to meet their own needs"[15]. Resilience, for the purpose of this thesis, can be defined as the "capacity of a system to absorb disturbance and reorganize while undergoing change to retain essentially the same function, structure, and identity"[16]. These are key considerations for future engineering systems considering ambitious global targets such as the United Nations Sustainable Development Goals (UN SDGs)[17]. In fact, important synergies can be found between flexibility, sustainability, and resilience. Improved allocation efficiency and adaptability to uncertainty are key drivers of flexibility, while representing essential objectives for sustainability and resilience- as captured graphically in Figure 1.1. Linking these objectives is particularly relevant in light of the UN SDGs, as designers and decision makers must look past traditional technical assessment to consider the holistic performance of a engineering system solution – a challenging task given the many tensions and computational limitations involved [18].



Figure 1.1: Flexibility, Sustainability and Resilience overlap

In light of the complexity and interactions associated among them, however, current design tools and frameworks are not well suited for the systematic integration of these important principles across an engineering system lifecycle. As a result, there is a substantial gap between their perceived and actual value under uncertainty[19]. This is driven by the inherent difficulty in predicting future scenarios, leading decision makers to often prioritize short-term financial targets or perform overly narrow evaluations where these concepts are not explicitly optimized. Based on these data and computational limitations, optimization of technical or economic metrics is normally prioritized in engineering systems, with limited scope or consideration of uncertainty in the early stages of the design decision making process[18]. Ultimately, this leads to reduced expected value across many multidisciplinary metrics[20], with systems highly vulnerable to the potential impacts of climate change[21].

Several frameworks for designing engineering systems based on resilience or sustainability have been developed in the past, but limited examples can be found integrating the two concepts**.** The two objectives are generally agreed to present significant overlap, but they can also be at tension with each other, and thus are not normally combined in the design and optimization stage - outside of specific applications. This may be due to locational factors, as countries where the risk of disaster is generally lower have prioritized sustainability with respect to resilience and vice versa, or computational limitations associated with a multi-objective evaluation[22-24]. Resilience is partially and implicitly integrated in several sustainability assessment design frameworks considering an inclusion of system ability to maintain performance as a result of climate change consequences and vulnerability[25-28]. Some promising attempts have also been made to incorporate sustainability into resilience assessment frameworks, particularly in the civil infrastructure or geo-technical engineering sectors where the link between the two is stronger [21,29-31]. The majority of the more quantitative based frameworks extend the life-cycle costs analysis to include environmental impacts of various designs and operational policies

[30,32-34]. A few qualitative systems design frameworks have also been developed but with limited application as they tend to be quite abstract and/or domain specific[21,35-37]. The integration of sustainability or resilience with an analysis of flexibility and/or real options for systems engineering design has been very limited thus far and no example is found holistically incorporating all three[38-41].

1.3 The Role of Flexibility Moving Forward

Flexibility in Design is a paradigm which is not yet fully exploited in the engineering community, based on several reasons. Decision makers often improperly estimate the added value it may bring to a project, while the costs are readily evident. Typical project evaluation approaches for engineering systems tend to be based on likely scenario predictions for major uncertainty drivers and Discounted Cash Flow (DCF) analysis, which is not able to account for the value of decision making over the project lifetime[42]. Nevertheless, both qualitative and quantitative tools exist to identify possible strategies and estimate their value. These are very important as they help bridge the gap between the perceived and actual value of embedding flexibility within a design, increasing the likelihood for decision makers to recognize their importance. Real options analysis (ROA) lies at foundation of Flexibility in Design, suggesting some general strategies such as abandonment, capacity expansion, altering of operating scale or staged capacity deployment- which are applicable across general infrastructure projects [43]. These methods have been shown to improve overall expected system performance in several previous studies across different engineering systems, typically by at least 10-30% [38,44-52]. In order to maintain reasonable computational expense, however, previous studies normally consider only one objective, while limiting the sources of uncertainty and flexibility decisions evaluated [53].

In light of the above, energy systems can be seen as both an important historical motivating example and potential impactful application area for these concepts. On one hand, transmission networks have usually been built via a cost minimization approach, with lowest required ratings to meet technical standards, and are thus highly vulnerable to a number of factors [54,55]. For instance, they are normally built above ground rather than below, leading to reduced reliability in the case of an extreme weather event. Similarly, they have often been planned with large fossil fuel generators supplying energy to cities which are quite far away, leading to very significant transmission losses and associated emissions. This creates a decreased reliability of service in case of any fault, as discussed above, as well as high sensitivity to input prices and operational requirements. Through the appropriate consideration of flexibility, however, it may be possible to invest more in the initial infrastructure to allow expansion of transmission capacity as needed in different sections of the grid. More sources of renewable energy and storage could thus be integrated over time based on network capacity and technical limitations. This allows an improved adaptability by system operators to respond to changing demand profiles and weather conditions – likely ultimately achieving decreased expected costs and overall enhanced resilience and sustainability.

Driven by a high degree of uncertainties experienced, several other opportunities arise for implementing flexibility concepts in the early design phases of future energy infrastructure investments. Globally over 940 million people still do not have access to electricity, which has important effects on issues such as education, economic development, productivity and health among others [56]. The cost of rural electrification via grid extension in low-income countries has become a major limiting factor in achieving energy access targets, and more modular capacity building may provide a better approach for investment capital limited settings [57]. The technical characteristics of these systems, as well as promising results from previous studies [57-60], suggest they can cost-effectively yield substantial improvements on sustainability and resilience metrics for achieving electrification compared to centralized deployment– given appropriate considerations of flexibility.

More advanced approaches have historically relied upon decision analysis, as well as Dynamic (DP) and Stochastic programming (SP). The value of flexibility in these methods is normally estimated as the difference in expected performance between the baseline and flexible design solutions. These tools, however, build upon a number of assumptions on path independence, knowledge of precise probability distributions, and market conditions such as arbitrage enforced pricing[43]. As such, they are usually unsuitable for implementation in an engineering setting, as decisions made during early design phases influence the techno-economic feasibility of various lifetime measures[61]. Furthermore, results are of limited relevance if realised uncertainties deviate from predicted scenarios.

A promising approach to analyze engineering systems for flexibility is to exploit a decision rule (DR) formulation. DRs are akin to *if-then-else* statements e.g., *if* demand exceeds a certain threshold, *then* expand capacity, *else* do nothing. An important benefit from this formulation is that it combines both physical and managerial aspects into an elegant and succinct formulation, making it rather intuitive to use by system operators and decision-makers. It can be used with advanced optimization techniques to identify the best system design configurations. DRs act like signposts, or triggering mechanisms, that help operators identify the conditions when it is best to exercise flexibility in operations. It is important, however, to carefully enable the flexibility to be used in operations via decision rules in early design activities, from an engineering standpoint. Several studies have shown the benefits of such approach to flexibility analysis, as compared to standard ROA based purely on DP or SP [62-64]. One issue with the approach, however, is that DRs are often generated by system designers based on a limited set of standard real option strategies. They are elicited through designers' expertise with a system, using for instance creativity techniques like brainstorming or prompting [49], and are not likely to completely explore the available design space. Thus, they may not include potential combinations, designs, timings, rules, and strategies to enhance the value of the system under uncertainty.

The strengths and weaknesses of current approaches are succinctly summarized in Figure 1.2 below. The graphic suggests that a data-driven perspective is likely to contribute substantially to the field,

allowing for a deeper exploration of the potential design solution space under uncertainty. The use of neural networks to capture approximate flexibility strategies, in particular, is compatible with a much greater dimensionality than possible with standard methods – as well as more realistic assumptions to engineering systems in practice. These advantages are key motivations for the objectives of this thesis, and they are explored in more detail in the following section.

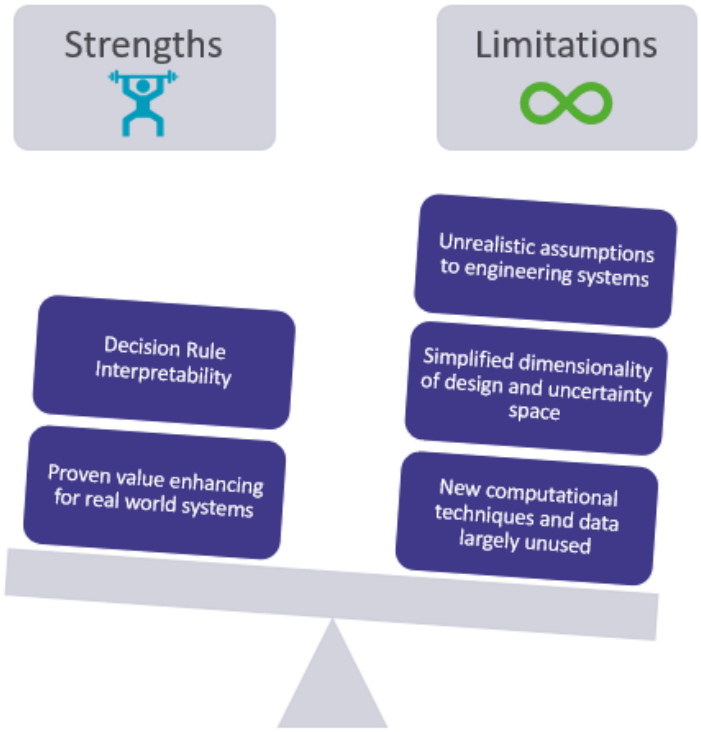


Figure 1.2: Analysis of trade-offs motivating the development of a new method in the field.

1.4 The Potential for Data-Driven Approaches and Machine Learning

Prior to the work presented in this thesis, flexibility design tools presented significant limitations on scope, applicability and integration into holistic systems engineering methods, which may potentially be addressed by the emergence of Machine Learning (ML) and data driven approaches. On one side, Flexibility in Design methods generally do not fully integrate or directly link sustainability and resilience objectives within their overall evaluative processes, partly to maintain problem tractability and computational feasibility [48,65]. These objectives present a significant amount of overlap for the reasons discussed in 1.2, and as such evaluating them together in the early stages of the project can improve performance and budget allocation efficiency. On the other side, they must usually be combined with some engineering creativity methods to generate flexibility strategies outside the standard real options framework. In fact, these methods are often focused on simply estimating the value of flexibility in a system, aiming to determine if it is a worthwhile investment.

Tractability and creativity limitations of standard flexibility in design methods, combined with the increased digitalization and emergence of newly available datasets, highlight the potential for ML in this field. By extracting useful patterns out of large datasets and increasing the ability to tackle high dimensional problems with neural network (NN) function approximation, the role of ML within engineering design offers great promise[66]. ML techniques may also be used to identify alternative system configurations, allowing a more systematic exploration of the design space data with a broader scope and through a combinatory evaluation not possible with human agents. Using ML can help to mitigate risks and uncertainty in the design process at lower computational costs than traditional approaches, yielding enhanced and adaptable infrastructure systems which are more resilient and economically viable in the face of climate change across a wide range of scenarios [66].

Reinforcement Learning (RL) in particular has shown great promise in ability to handle time-dependent control tasks and outperform alternative state of the art methods which are outside the scope of traditional ML[67,68]. This is partly due to enhanced ability to incorporate physics models and expert knowledge of design space compared to supervised learning methods[69]. Recent notable industry applications range from data centre energy management[70] or search engine optimization[71] by Google, to general purpose Large Language Models by OpenAI[72] or Meta[73], among others. Their results point to the potential for addressing very complex problems through this approach. Furthermore, RL is very well suited to the sequential-decision making nature of flexibility analysis under uncertainty, able to learn effective strategies through a very large number of interactions[67]. A deep reinforcement learning (DRL) formulation to flexibility analysis enables an agent (i.e., the system operator) to analyze different strategies based on a heuristic learning process. Accordingly, it can *learn* the best actions under different conditions through statistical training over many samples. Consequently, DRL may help identify new dynamic adaptation strategies from the data, as opposed to relying solely on those elicited from standard real options and flexibility in design methods. It enables uncovering different combinations, timings, and flexible designs that may not be considered explicitly by system designers, thereby expanding and improving the design space exploration process.

1.5 Motivation Summary

Table 1 below presents a general summary of proposed methodological gap tackled by this thesis, based on the issues discussed thus far in this chapter . The motivation and contribution are discussed in more detail in 1.6 below. They are based on a thorough review of the literature presented in Chapter 2 where each field and corresponding relevant works are presented in more detail in their respective sections. Nonetheless, this high-level summary is shown here to provide some early intuition on the context and overall objectives guiding the methodological developments presented in this thesis.

Table 1: Overview of research motivation driving this thesis, with the works cited further explored and discussed in Chapter 2.

Systems Design Methodology	Sustainability and Resilience Enhancement	Systematic exploration of design space	Able to tackle higher dimensional problems	Usability by decision makers	Integration into design framework	Exploitation of value enhancing flexibility	Dynamic and systematic risk tolerance integration
Real options analysis[11,39,40,61,74,75]	☑	☒	☒	☑	☑	☑	☒
Dynamic Programming/ Decision Analysis[53,76-79]	☒	☒	☒	☑	☑	☑	☒
Decision Rules[62,80-82]	☒	☒	☑	☑	☑	☑	☒
Reinforcement Learning[66,83-87]	☑	☑	☑	☒	☑	☒	☒
Proposed methodological gap for this Thesis	☑	☑	☑	☑	☑	☑	☑

1.6 Research Objectives and Contributions

1.6.1 Summary and Research Objectives

The paradigms of flexibility, sustainability and resilience in engineering design are yet to be integrated in a comprehensive manner as part of a framework which can systematically explore the design space of a system to capture their value, particularly under uncertainty. Flexibility evaluation tools based on ROA can provide useful quantitative metrics, but are limited by unrealistic assumptions and the curse of dimensionality, as discussed in 1.3. DRs, while, more intuitive in nature, normally require potential flexibility strategies to be pre-selected and must be combined with creativity tools for end-to-end solutions. Thus, they are potentially missing out on optimal alternatives that may exist from available data, or in combinations that may not be intuitive to consider for system designers outside of generic strategies. ML techniques are still relatively underexploited in this area

but show great potential to address some of the data and computational limitations imposed by existing methods in Flexibility in Design. ML may be used to combine these methods in different ways by exploring the data more systematically, leading to more complementary solution procedures to analyse flexibility under uncertainty.

DRL offers the most promising methodological solution in the field of complex systems engineering design and planning. The sequential decision-making nature of the RL problem may be an improvement to DP as well as complementary to other evaluation techniques, with more flexibility on assumptions taken and dimensionality of the problem. Particularly interesting is considering the conceptual integration of RL with the DR approach from ROA, as it provides a more intuitive and simpler means to understand the internal analysis performed by the learning agent, while stimulating creativity on alternative strategies through a deeper exploration of the solution space. The theoretical common ground shared with standard methods allows a more straightforward solution procedure to be created based on current common flexibility evaluation methods. This helps to address some of the concerns on "black-box" calculations performed by ML techniques, while leveraging the computational power of neural networks to provide a usable set of guidelines. The overarching research question motivating the work presented in this thesis is thus as follows:

How can a data-driven based approach to analyse flexibility in engineering systems design under uncertainty best complement existing methods in the field?

This questions is of great interest to the scientific and industrial communities in ML, Systems Design, and Renewable Energy in general as the potential of DRL for such a task remains unexplored. In order to progressively validate this hypothesis, the broader topic is divided into a set of more specific research questions. Together, they highlight the originality and contributions of this thesis to the field, and rigorously evaluate the proposed approach for different criteria and potential applications. The chapter where each research question is primarily addressed is highlighted in bold.

- RQ1: How can ML and other data-driven methods add value across the different phases of the Flexibility in Design Framework? **(Chapter 3)**
- RQ2: Can DRL be used to analyse flexibility in engineering systems design under uncertainty as part of a standardized process? **(Chapter 3)**
- RQ3: Is it possible for DRL to outperform existing methods in Flexibility in Design on a well-known proof of concept case study? **(Chapter 4)**
- RQ4: Does the proposed data-driven approach enhance expected value through flexibility even for more complex and higher dimensional engineering systems design problems? **(Chapter 5)**

- RQ5: What are the computational limitations of DRL in terms of case study complexity and can it handle design problems that have never been tackled with alternative methods? **(Chapter 6)**
- RQ6: What novel techniques may be explored to improve the interpretability of the proposed approach and enhance acceptability in practice? **(Chapter 4, Chapter 5, Chapter 6)**
- RQ7: What other valuable application areas are possible with the proposed approach outside of direct flexibility evaluations? **(Chapter 7)**
- RQ8: How can risk profiles be integrated through a DRL formulation in a more dynamic and systematic way than currently possible with alternative methods? **(Chapter 8)**

1.6.2 Challenges

This thesis presents the development of a novel approach to analyse flexibility in engineering systems based on a data-driven perspective. The framework proposed allows for a more realistic evaluation of complex engineering systems in the early design phases, leveraging flexibility concepts to enhance sustainability and resilience objectives - while still capturing the financial realities of these investments. Based on the analysis of available literature, it represents a significant advancement in the state-of-the-art, substantially increasing its scope and applicability across different decision-makers and designers of engineering systems. The main objective of this thesis was to develop a novel computational data-driven approach to analyse flexibility in complex engineering systems under uncertainty more deeply, aiming to complement existing methods. The proposed approach ought to be compatible for the planning of different engineering systems, although the energy space is deemed of particular interest due to its abundant data availability, potential value of flexibility, and impact on sustainable development. The primary challenges and improvement in targeted properties for the novel approach proposed in this thesis are further summarised below, all as compared to current state of the art methods, as well as where in the document they are primarily addressed.

1. *Effectiveness:* the data-driven methodology employed should yield more optimal or better expected performance solutions under uncertainty than alternative approaches.
    - This is addressed in all chapters of the thesis, as any solution obtained via the proposed approach is benchmarked with relevant existing methods.
    - RQ1, RQ2, and RQ3 are the primary drivers for this contribution objective.
2. *Dimensionality*: the proposed approach should be able to handle more sources of uncertainty, flexibility decisions, and objectives.
    - This is progressively addressed from **Chapter 4** to **Chapter 7**, as each case study tackled presents higher dimensionality and uncertainty space than the former, although it is a point of focus throughout the work.
    - RQ3, RQ4, and RQ5 are the primary drivers for this contribution objective.

3. *Creativity:* the framework and methodology should yield alternative timings, decisions, and strategies through a data-driven exploration than with existing methods.
    - Several approaches are explored to exploit the additional strategy creativity resulting from a deeper design space exploration, mainly from **Chapter 4** to **Chapter 6.**
    - RQ6 is the primary driver for this contribution objective.
4. *Adaptability:* the method should be easily adaptable to different kinds of engineering systems as well as to the prioritization of different objectives.
    - The case studies tackled from **Chapter 4** to **Chapter 6** represent very different types and scales of engineering systems, showing the adaptability of the proposed approach. The work in **Chapter 7** also shows it can be extended outside of direct engineering design to policy-tool formulation objectives.
    - RQ3, RQ4, RQ5, and RQ7 are the primary drivers for this contribution objective.
5. *Acceptability*: decision makers should be willing to implement the developed framework and methods in practice, particularly in light of general resistance to "black-box" optimization.
    - This is a common theme throughout the thesis and a point which was stressed from the very beginning of the PhD project. Different methods are explored across the chapters, providing several interesting avenues for interpreting outputs from the DRL models. **Chapter 4** to **Chapter 6** each showcase a different approach for acceptability and interpretability under risk-neutral objectives, while **Chapter 8** introduces a distributional perspective for improved risk tolerance integration.
    - RQ6 and RQ8 are the primary drivers for this contribution objective.

1.7 Thesis Structure and Overview

The structure of this thesis is generally summarised by a description of each chapter below.

Chapter 2 presents an overview of the background and relevant literature to the objectives of this thesis. It explores in further depth the field of Flexibility in Design and available computational methods, including their advantages and limitations. This is followed by a brief general review of ML and its applications in engineering design, with a more in-depth presentation of the mathematical, theoretical, and practical background of RL techniques, which are most relevant to this work. A brief review of previous attempts at designing for sustainability and resilience via alternative methods is also introduced.

Chapter 3 proposes the novel framework and procedure for a data-driven Flexibility in Design method. The role of different ML methods is explored across phases of the design framework, although the focus is on the development of DRL flexibility analysis. This includes how to identify engineering systems where this may be particularly valuable, the problem formulation sequence, and

the methodological procedure employed to feasibly reach a solution. It summarizes the core methodological advances proposed and analyses them from a holistic implementation perspective.

Chapter 4 firstly validates how the proposed framework may be applied on a simpler proof-of-concept infrastructure system under uncertainty, which has been repeatedly used as a benchmark in Flexibility in Design. Results are extensively compared to state of the art, primarily along the desired properties discussed in 1.6. The proposed method is shown to significantly outperform alternatives. Limitations and opportunities for further research are also discussed based on these outputs.

Chapter 5 expands on the findings from Chapter 4 by tackling the significantly more complex case study of the design of a Waste-to-Energy (WTE) plant network commissioned by the city of Singapore. As opposed to the proof of concept from Chapter 4, several more sources of uncertainty are introduced, and flexibility decisions are expanded to include a spatial component across different sectors of the city. This also marks a shift in the algorithmic approach implemented, in line with the general literature on DRL discussed in Chapter 2.

Chapter 6 further stresses the feasibility of the proposed approach by presenting the first ever design and planning optimization of a mobile energy supply system for nomadic communities in Mongolia. This case study is of much higher dimensionality than any previously evaluated with Flexibility in Design methods, showcasing important results across economic performance, sustainability, and resilience. A sensitivity analysis on the impact of the potential development of market-support instruments is also initially explored here.

Chapter 7, originally inspired by the policy-tool sensitivity analysis presented in Chapter 6, focuses on the development of a novel computational framework for the data-driven formulation of policy-making tools under uncertainty. This novel approach is also based on DRL but is adapted to represent governmental and institutional perspectives. The case study tackled is primarily based on that detailed in Chapter 6, although potential policy cases are refined to be most relevant to national targets. The results show how this kind of analysis can contribute to enabling energy access for rural communities worldwide, particularly those following a migratory lifestyle.

Chapter 8 more closely examines the implementation of this method in practice, starting with a consideration of the parameters and constraints used to formulate the problem. In order to better captures decision makers differing risk profiles, a distributional approach to DRL is implemented to optimise metrics which are more closely related to the risk-tolerance of a particular investment. This is only explored on the proof-of-concept case-study from Chapter 4 due to the significant computational expense associated, and as risk-profile integration was not the primary motivation for the proposed approach. The results are very promising, providing the desired distributional shift, and outlining a blueprint for implementation on other case studies.

Chapter 9 concludes the work presented in this thesis, leading to several future research possibilities to expand and improve on the contributions provided in related areas. It presents a more holistic evaluation of the outputs of this thesis in practice, based on the findings from all the above sections. Trade-offs with alternative state of the art methods are discussed along issues with reward metric selection, intuition for algorithmic performance, implications for generalizability under uncertainty, and overall applicability in industry.

1.8 Publications and Collaborations

Some of the work presented in this thesis was also included in the following peer-reviewed journal and conference publications and conference proceedings. They are listed in chronological order below as well as which chapters they contribute to, respectively.

- Caputo, C. and Cardin, M-A. 2020. A machine learning based framework for the design and evaluation of flexible, sustainable, and resilient energy systems under uncertainty. Supergen Energy Networks Hub Risk Day, March 4th, Glasgow
    - **Chapter 3**
- Caputo, C. and Cardin, M-A. 2020. An Approach to Analyze Flexibility in Engineering Systems Design based on Reinforcement Learning and Decision Rules. Proceedings of the 2021 IISE Annual Conference, May 25th, Virtual.
    - **Chapter 3, Chapter 4**
- Caputo, C., & Cardin, M. (2021). The role of Machine Learning for Flexibility and Real Options Analysis in Engineering Systems Design. Proceedings of the Design Society, 1, 3121-3130. https://doi.org/10.1017/pds.2021.573
    - **Chapter 3, Chapter 8**
- Caputo, C., and Cardin, M. (September 21, 2021). "Analyzing Real Options and Flexibility in Engineering Systems Design Using Decision Rules and Deep Reinforcement Learning." ASME. J. Mech. Des. February 2022; 144(2): 021705. https://doi.org/10.1115/1.4052299
    - Special issue on “AI for Engineering Design”
        - https://asmedigitalcollection.asme.org/mechanicaldesign/article/144/2/020301/1128870/Special-Issue-Artificial-Intelligence-and
    - **Chapter 5**
- Caputo, C. et.al (2022). “Energy System Evolution Strategies for Mobile Micro-grids through Deep Reinforcement Learning” AI Futures Conference, June 14th, London.
    - **Chapter 6**
- Caputo C, Cardin M-A, Korre A, et al., 2022, Energy System Evolution Strategies for Mobile Micro-grids using Deep Reinforcement Learning Flexibility Analysis, 32nd European Conference on Operational Research (EURO 2022), Espoo, Finland

    - **Chapter 6**
- Ge P, Caputo C, Teng F, et al., 2022, A Wireless-Assisted Hierarchical Framework to Accommodate Mobile Energy Resources, IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Singapore
    - **Chapter 6, Chapter 7**
- Cesare Caputo, Michel-Alexandre Cardin, Pudong Ge, Fei Teng, Anna Korre, Ehecatl Antonio del Rio Chanona, Design and planning of flexible mobile Micro-Grids using Deep Reinforcement Learning, Applied Energy, Volume 335, 2023,120707,ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2023.120707.
    - **Chapter 6**
- Cesare Caputo, Pudong Ge, Michel-Alexandre Cardin, Pudong Ge, Fei Teng, Anna Korre, Ehecatl Antonio del Rio Chanona, Enabling Nomadic Energy Access via Plug and Play control and Deep Reinforcement Learning, Submitted to Nature Energy (*under review*)
    - **Chapter 7**
- Cesare Caputo, Michel-Alexandre Cardin, Ehecatl Antonio del Rio Chanona. Risk Sensitive Flexibility Analysis of Engineering Systems under uncertainty via Distributional Deep Reinforcement Learning. To be submitted to Nature Machine Intelligence (*in progress*)
    - **Chapter 8**

Outside of the above publications, I helped supervised many interesting MSc. projects with relevance to the overall objectives of the thesis. The main authors and project titles are summarised here, and while not all of them could be integrated the thesis directly, they provided crucial contributions and we are thankful to have had the opportunity to collaborate on these topics. In cases where material from collaboration is used (only applicable to Chapter 6 and Chapter 8), it is explicitly stated and appropriately credited.

- Joshua Anderson. “Design and analysis of flexible multi-layer staged deployment for satellite mega-constellations under demand uncertainty”. Department of Design Engineering.
    - Primary supervisor: Dr. Michel-Alexandre Cardin
- Jingui Wu. “The impact of time step resolution on Deep Reinforcement Learning for Engineering Systems Design and Planning”. Department of Computing.
    - Primary supervisors: Dr. Michel-Alexandre Cardin, Dr. Alessandra Russo
- Kedawat Ayush, “An empirical study of trade-offs between constraint violation and expected performance through Deep Reinforcement Learning Flexibility Analysis on engineering systems under uncertainty”. Department of Chemical Engineering

    - Primary supervisors: Dr. Ehecatl Antonio del Rio Chanona, Dr. Michel-Alexandre Cardin
- Alexander Gibson. “Designing an Intuitive Decision-Support System for the Operation of Future Energy Systems”. Department of Design Engineering.
    - Primary supervisor: Dr. Michel-Alexandre Cardin

Chapter 2

2 Background and Related Work

This chapter begins with a fundamental review of the field of Flexibility in Design. A brief background on ML is presented, with a stronger emphasis on DRL theory and application. It should be noted that this is not comprehensive given the varied fields covered. A more general overview of Flexibility in Design can be found in[61] while the foundational material on RL is presented in full by [67].

2.1 Flexibility in Design

2.1.1 Origins and Real Options

A flexible design provides the potential for changeability of a system under different and unpredictable conditions, through inclusion of some potential strategy into the system itself. One of the most comprehensive reviews of flexibility in engineering design is that found in [61], which builds upon the definition of flexibility as "enabling a system to change easily in the face of uncertainty" from a technical and technological standpoint [48]. This will be the definition used in this thesis, although the aim will be to also expand it to include social and environmental perspectives, when possible, in order to allow compatibility with the UN SDGs. Therefore, flexibility comes from inclusions in the original technical design of some measures to allow change, as well as the ability to actually enact on this at some point during the project and yield the desired modifications [61]. The ability to enact on this flexibility measure is akin to the definition of a "real option"[74].

A real option "in" the system is composed of the specific factors in design and management which allow the enhancement of flexibility during system operation, or exercising the real option "on" the system [61,88,89]. Real options theory stems from financial literature theory, and thus gives an array of quantitative methods to objectively evaluate the value of flexibility under different scenarios, which allows better comparison of alternative design concepts. The distributional shift in expected performance which these tools seek to produce is shown in Figure 2.1, reducing downside exposure and better positioning for upside potential. This is captured in the tail ends of the distribution (i.e. P5 and P95) which, while perhaps less likely to happen, can cause the largest deviations from predicted performance – although still looking for an overall improvement in expected value. Real option theory, however, is limited in that it can only suggest generic design alternatives that are able to enhance flexibility, with limited consideration of the holistic system, thus must be combined with Flexibility in Design methodologies to yield complete, better performing solutions[48,61].

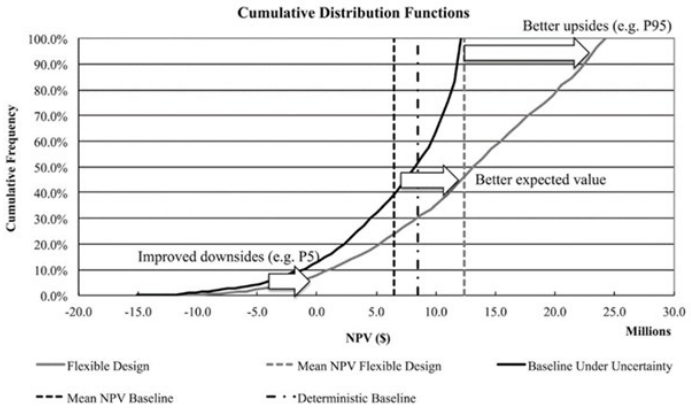


Figure 2.1: Example cumulative distribution functions for conceptual baseline and flexible systems design concepts, using NPV as lifecycle performance metric[61].

2.1.2 Definition and Framework

An important distinction between Real Options and Flexibility in Design is that the former focuses on quantifying the value of flexibility – effectively aiming to price real options – while the latter focuses on developing methods and procedures to embed flexibility in engineering systems design, as a value-enhancing mechanism. Both fields go hand in hand, since one enables quantification of the value of flexibility that is designed early on in each system. Flexibility in Design relies on value quantification in a similar fashion as in real options theory, but more as a mechanism to rank order possible design alternatives to support the decision-making process, and less to find the right "price" for the real options. It adapts the theory of real options to accommodate the needs of industry practice[10].

Previous works have generally attempted at structuring designing for flexibility through a sequential process workflow in order to provide standardized and reproducible procedure[48,90,91]. In the most recent framework and taxonomy presented in [61], five main phases are recognized for a flexible design process, as captured in Figure 2.2. Baseline design in 1), conducted with traditional optimisation methods, is necessary to provide a point of comparison and understand relative performance of output designs. Procedures in this phase can help to generate new design solutions [92], enrich current information on available solutions like axiomatic design [93] or help navigate the design process, such as in concurrent design[61]. Concept-Knowledge theory, function based failure analysis and Bayesian architecture generation are some further methods which can be used to conceptually navigate the design space and yield a set of baseline designs essential for flexibility analysis[94,95]. Uncertainty recognition in 2) is essential and can be a defining factor in the design of a project under stochastic conditions, as estimates from this phase may skew results towards largely unrealistic conditions or ill-prepare designs for realized scenarios[6]. Practical approaches to model them then can include stochastic simulation-based diffusion models, binomial lattice, decision trees and scenario planning. The selection of input parameters, assumptions and mathematical approaches can have a significant

effect on predictive accuracy of these models, with different techniques better suited to handle different types of uncertainties or design problems, although with all presenting some important limitations. Phase 3) then involves generating performance-enhancing system design concepts leveraging flexibility, followed by 4) design space exploration and optimization. The framework is embedded within a holistic phase 5) to enable seamlessly the design process among relevant stakeholders. Out of these phases, this thesis is deemed to make the greatest impact on Concept Generation and Design Space Exploration, as further detailed below.

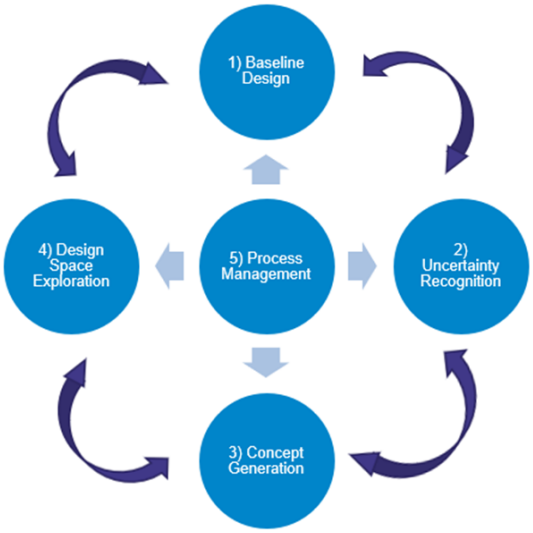


Figure 2.2: Flexibility in Design Framework

2.1.3 Creativity and Concept Generation

Creativity is essential for systems engineering, helping to identify alternative solutions and strategies to solve the specific design problem. Its integration within overall flexibility design methods, however, remains fairly limited [96,97]. Creativity in design must be looked at within the bounds of a product's or systems functional requirements, creating complex trade-offs between stimulus for novelty and adherence to original objectives[98,99].

The tools available for concept generation, with a particular focus on enhancing system flexibility, can be quite varied ranging from qualitative to computational approaches and can be further split into strategy generation and enabler identification [61]. These two parts are complementary as they both must be present to some degree to allow for the exercising of system flexibility under uncertainty on a particular system. During the strategy generation phase, there are a number of generic real options strategies which can be used to guide creativity in this phase such as : (1) defer capital investment until favourable market conditions arise, (2) stage or phase asset deployment strategically over time instead of deploying all capacity at once, (3) alter operating scale by expanding or contracting output production capacity, (4) abandon a project doomed to fail and resell assets at salvage value, (5) switch

production inputs and/or outputs to accommodate different markets or missions, (6) invest in R&D to capitalize on future technology and cash flows if the initial investment is successful, and/or (7) combine the above [43]. Creativity then stems from the need to understand how to integrate these approaches into the specific system hand. The Integrated Real Options Framework formalized this latter step by mapping potential measures to uncertainty drivers. The logical multiple domain matrix is an extension of this work to generate concepts for flexibility via a more holistic analysis of interactions among domains[100].

In terms of enabler identification, a variety of approaches are also present. An initial starting point may be product or system safeguards, connecting modularity to degree of uncertainty as shown by[101] . Another commonly used tool in this phase is the Design Structure Matrix (DSM). The DSM consists of listing system components in the rows and columns of a square matrix in order to map out dependencies and relationships among them. It is often used in combination with Change Propagation Analysis (CPA) to understand the response of different design variables to changes in design or functional requirements and identify candidate areas for value enhancing flexibility[102] . The method has been previously successfully extended to include non-technical domains (e.g., Engineering Systems Matrix from[103]) as well as Bayesian network analysis[104], however the effort and level of detail required to build this representation can be limiting. A similar analytical approach is followed when using the Theory of Inventive Problem Solving in the context of designing for change to identify candidate areas for modularization[105]. The Filtered outdegree method assigns a value to the perceived degree of changeability of a system, with possible end states filtered based on a transition cost function and CPA [106]. While this method is most useful for measuring flexibility in a system, it can also help identify enablers through embedding of real options when combined with CPA or other techniques, essentially serving to decrease the cost of transition in the outdegree. Explicit design variable evaluation is another option, consisting of identifying the main design variables involved in the system and quantitatively evaluating their impact on lifetime performance, helping to identify the most impactful changes[61,107]. The key trend, however, is that these tools are lacking a way to systematically explore the available data in order to enhance sustainability and resilience objectives through flexibility, likely leading to underexploited value enhancing strategies for the system.

2.1.4 Design Space Exploration

Following the generation of flexible design concepts in 3), potential strategies must be evaluated on a comparative basis to understand what the best performing solutions may be in the system at hand. Dynamic (DP) and stochastic programming (SP) are alternative solution methods to ROA for flexibility evaluation in decision-making problems under uncertainty. They seek to compute an optimal policy in form of maximization of the recursive Bellman reward equation, normally as part of decision analysis by dividing a complex problem into a sequence of simpler subproblems [67,108]. DP is better suited to handle infinite horizons problems or ones with large number of stages compared to SP,

however both are limited by data inefficiency, memory storage requirements and dimensionality [109]. Binomial lattice methods are similar conceptually to DP, other than the fact that uncertainty either goes up or down following each stage. The value of flexibility in these methods is normally estimated as the difference in expected performance between the baseline and flexible design solutions. These methods, however, present a number of assumptions on path independence, knowledge of precise probability distributions and market conditions such as arbitrage enforced pricing which make it generally unsuitable for implementation in an engineering design setting[89,110]. Path independence assumes that the value of a system at a particular state or node is the same regardless of the sequence of operations preceding it to reduce the number of possible outcomes[111]. This is often not the case in engineering systems, particularly as decisions made during early design phases can affect technical and economic feasibility of various measures throughout the project. Furthermore, results are of limited relevance if realised uncertainties deviate from model scenarios[53]. Other approaches have been proposed in recent years to help address these limitations such as screening methods[112] or design catalogues[113]. Drawbacks of these techniques, however, may include model resolution loss, limited guarantee of solution optimality, and difficulties with estimating aggregate preferences to calculate expected utility in a design context[61].

2.1.5 Decision Rule Formulation

One issue with the above approaches is that the decision to exercise flexibility is difficult to use in real-world operations. Among others, the folding back process in DP (i.e. starting from the end and folding back to initial time) is often criticized as not intuitive and difficult to implement in practice [62]. The decision rule is essentially based on Bellman's expected reward maximization, and therefore does not provide much flexibility in terms of implementing other rules. It may also prove challenging in operations to identify the corresponding system state in a decision tree or lattice and determine the next best decision.

To address these issues, a decision rules approach to flexibility analysis was proposed recently, and slowly taking precedence in the field [11,62]. Flexibility Decision Rules (FDR) are typically employed in stochastic programming and robust optimization to alleviate intractability in the computational problems created. They include condition-go (akin to the *if-then-else* formulation), zero order, linear, and safety-first rules [114]. In the context of flexibility analysis, decision rules are not just mathematical tricks, they take on a full managerial meaning, giving intuitive policies as to when flexibility should be exercised, and how. An example FDR formulation may be that "*if* demand reaches threshold α, *then* expand capacity by β, *else* do nothing". Given that the decision may be evaluated a regular time intervals, the problem can be modeled as a multi-stage stochastic programming or robust optimization problem, to identify the best values for decision rule variable α and design variable β over the problem time horizon [11]. This formulation leads to intuitive policies readily applicable in operations and connects tightly what is done to the physical design to enable the flexibility early on to how it is

managed in operations. The approach has been shown previously to quantify the value of flexibility in a similar manner as standard real option approaches [62,80-82].

A decision rule formulation is embedded within the definition of a *flexible system design concept*, which comprises both a *strategy* and an *enabler* [61]. During early conceptual design activities, designers can choose a strategy to deal flexibly with certain uncertainty sources (e.g., price, demand) by considering standard real options (e.g., abandonment, expansion, etc.). They must also identify how to enable flexibility in the early design phases, which includes considering both physical design aspects and a policy to determine when it is best to exercise the flexibility during operations (i.e., the decision rule). Historical examples, however, show that there is no "one-size-fits-all" solution to design flexible systems. Different systems face different uncertainty sources and require different design configurations. Hence, a combination of creativity, guidance, and expertise is needed to identify the best strategies and enablers to embed valuable flexibility in a particular system[10].

The DR method is, however, fairly static in nature in the sense that the DR are normally determined at project start and not adjusted thereafter. Thus, while it allows some degree of flexibility as it permits system response to realized scenarios and trends, it is incomplete in that it does not allow more adaptive decision making and optimization of performance, only making limited use of observations and reducing potential flexibility value from embedded measures [115].

2.2 Machine Learning and Deep Learning

2.2.1 Background

ML is the general term describing the field of study of using algorithms and computational power to extract usable models and patterns out of empirical data, although definitions vary widely [116]. Generally, it can be categorized into supervised and unsupervised learning and there is an abundance of potential algorithms available[116,117]. Supervised learning provides labelled datasets to predict their relationship and is primarily used in classification and regression. Unsupervised learning instead seeks to find patterns and distributions from unlabelled data and can be used in clustering or feature extraction applications, among others. The recent resurgence of ML can be primarily attributed to the breakthrough in "Deep Learning" with the implementation of a hierarchical feature extraction through neural networks (NNs) [118].

This layer by layer learning structure within a network allows for much more complex feature extraction than possible when using raw inputs directly, which combined with advances in data availability and computing power have allowed very important results to come out of this field [119]. The universal approximation theorem, in fact, yields that any continuous function can be estimated with various levels of precision by a neural network through changing the number of neurons with at least one hidden layer[120]. The input to each layer after the first is normally composed of a weighted sum from the previous one, transformed via a generally nonlinear activation function to link the

information between different layers and perform the calculations[118,119]. It may consist of several layers – hence the name "deep". Throughout training, these weights between layers are thus updated to optimize some loss function defined initially, normally using gradient information across the layers[118,119,121]. The scalar values captured in each layer are typically called units or neurons. The intuition behind the value of neural networks is their ability to decompose higher lever features into smaller ones found in raw inputs, combining distributed representation with end to end training[68,119,120]. The complexity and *depth* of the network architectures is based on trade-offs among granularity of the approximated representation, computational limitations, and ease of training.

2.2.2 Optimization and Key Features

The optimization of NN parameters is generally based on variations of gradient-descent optimization. Firstly, a batch of inputs is used to implement a forward pass through the network layers, generating the corresponding desired outputs. A loss function is then computed, with a corresponding gradient across all network parameters, in the form of partial derivatives. Iterating from the output layer, the chain rule allows for the sequential calculation of gradients in a backward pass, all the way to the input layer. This process, known as backpropagation, is fundamental to the efficient training of NNs. The computation of derivatives, in particular, is performed through automatic differentiation, which is able to estimate them even in the lack of an analytical or user-defined derivative functions[120].

Several variations of vanilla gradient descent now exist which can speed up convergence. They include techniques such as momentum, to exploit gradient information from previous updates, or the computation of second order derivatives to based network updated on more information. Their purpose is generally to address noisy gradients, reduce chances of getting stuck in local minima, and ultimately improve training efficiency or wall-clock time. Stochastic Gradient Descent (SGD)[122], AdaGrad[123], Adam[124], and RMSprop[125] are some of the most commonly implemented approaches, although the field continues to evolve.

Rather than using the entire dataset, computations are typically performed in smaller batches, helping to address memory limitations, and exploiting modern computational parallelization possibilities. As deeper networks are implemented, *vanishing gradients* (increasingly smaller values) may become problematic, particularly depending on the activation functions used. These are typically non-linear to allow for enhanced expressivity, with ReLU (rectified linear unit), sigmoid (logistic), and tanh (hyperbolic tangent) among some of the most popular. Depending on how the different layers are connected, they may referred to as multilayer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN) or long short-term memory network (LSTM), with each architecture appropriate for specific tasks [118,119]. The final output layer is normally characterized by a probability distribution, generally exploiting either a logistic function or softmax for targeted discrete outputs. Continuous domains, on the other hand, tend to implement a multi-variate Gaussian

distribution to generalize to higher dimensionality of problems tackled[119] – although other alternatives are possible.

2.2.3 Machine Learning for Engineering Design

The potential of ML and data driven methodologies to enhance the systems design process is of great interest, and examples of applications within the engineering design space are abundant, but they tend to be fairly limited in scope[126]. There has been great success working with and extracting patterns out of time series data and increasing prediction and simulation accuracy[66]. These models can handle much larger data sets than would be feasible with conventional approaches to reproduce the behaviour of complex non-linear systems, often with much better computational efficiency than analytical solutions. In the engineering space, this may be useful to predict operating conditions for a system and better characterize the technical requirements to meet project objectives throughout its lifetime. In the energy space, the increasing amount of data has created a number of potential and successful use cases, particularly in prediction and scheduling for improved uncertainty recognition [117,127]. In the design space exploration and optimization phase, classification algorithms can be used to identify appropriate regions of feasible design space among a number of candidate solutions. This is normally done through finding the values for some specific variables that satisfy a number of constraints while potentially optimizing an objective function at the same time, which is an example of an inverse design problem [128]. These approaches have been particularly useful when trying to map sub-design problems at different levels of the system[129]. Mapping the boundaries of the feasible design space can help guide engineers on more efficient exploration of the space while limiting the need for computationally expensive simulations at every stage. This has been important in reliability-based design to optimize the analysis efficiency on high dimensional problems such as in additive manufacturing, materials design, aircraft wings, communication systems and structural design among others [129,130]. Image processing has been particularly important in optimization tasks in computational fluid dynamics and process engineering settings to optimize and efficiently evaluate features such as turbulence or mixing ratio in reactions with relatively lower level inputs [131].

Support Vector Machines, Random forests, Gaussian Naive Bayes and ANNs are some of the prevalent ML techniques previously used for engineering design, comparatively evaluated by[129] on several different application problems. The findings from their study suggest that while no approach is universally dominant, these techniques hold significant potential in helping to facilitate and speed up design space exploration across various disciplines. The success stories of these applications have, however, been heavily dependent on the quality and quantity of data fed to the network, which tends to require significant time-intensive data pre-processing to avoid issues such as bias or overfitting, among others. The majority of applications so far have, therefore, focused on increased accuracy of prediction, scenarios, surrogate modelling, and classification as inputs into design but have yet to be integrated into an end-to-end design process to stimulate creativity about alternative flexibilities or

strategies. On the other hand, the creativity assessment work by [84] is interesting, where ML is used to automatically rank a large number of designs collected via a crowdsourcing platform on novelty and feasibility with 3 separate ANN, which showed some of the creative limitation of this approach. There has been some concerns on the practical applicability of these methods for decision makers due to the "black-box" computation process, whereby there can be millions of operations happening between input and output of which limited people can claim a complete understanding of [85].

Finally, in the review by[132] on the intersection between ML and computational creativity, the authors argue that while the most obvious use of ML would be as a means of increasing exploratory creativity, there are a number of other potential applications. For instance, classification or regression could be used to learn the evaluative or objective function to be used in the design optimization procedure, thus developing a new way to analyse the system at hand. Another application could be using predictive models to generate new designs or solutions based on incomplete ones (combinational creativity), which could be extended to learning the generative model directly (exploratory creativity). Data mining is also argued to be useful for discovering nonobvious links among system components and the resulting implications for design.

2.3 Deep Reinforcement Learning

2.3.1 Background and Notation

2.3.1.1 Markov Decision Processes

The mathematics and theory behind DR formulation and solution procedures are closely related to a Markov Decision Process (MDP), which characterizes the typical RL problem. MDP is the standard learning framework for sequential interactions between an agent and its environment. MDPs are an extension of Markov chains, augmented by actions selection possibilities with the aim of reward maximization[133]. As such, they are typically characterized by an action space $\mathcal{A}$, a state space $\mathcal{S}$, a state transition probability distribution $\mathcal{P}$, and a reward function $\mathcal{R}$. Generally, at each time step t, the agent finds itself at state s^t, and selects an action a^t following a decision-making policy π. This policy maps different state observations to probabilities of action selection such that $\pi(\, s^t | \, a^t)$. As a result of taking the action, the agent transitions to a successive state $\mathrm{s}^{\mathrm{t+1}}$ receiving a corresponding scalar reward $\mathrm{r}^{\mathrm{t+1}}$, which are governed by the environment's dynamics and reward function $\mathcal{R}(\mathrm{s}^{\mathrm{t}}, \mathrm{a}^{\mathrm{t}})$, respectively [67]. The objective of the agent is to maximize *discounted* expected reward $G_t = \sum_{k=0}^{\infty} \gamma^k \, R_{t+k}$ from each state s^t. A discount factor $\gamma \in [0,1)$ serves to bound the sum of rewards in infinite episodes, while also prioritizing short term rewards over long term ones, akin to the approach used in financial modeling for NPV calculations [134]. The sequential decision-making process which captures the reinforcement learning problem is thus shown in Figure 2.3.

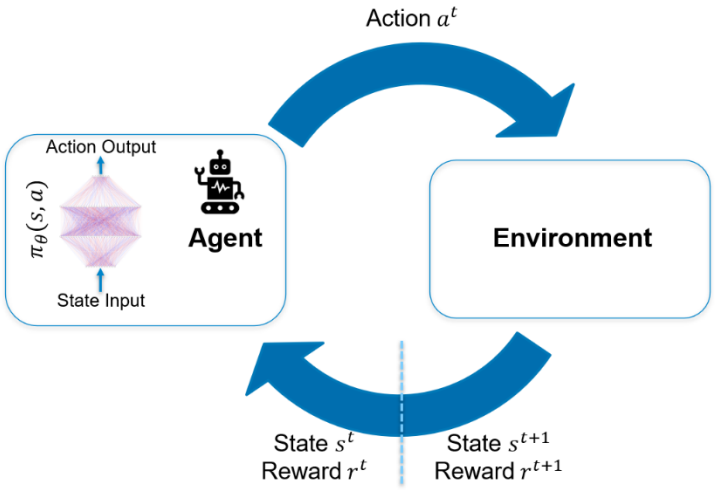


Figure 2.3: DRL agent-environment sequential interaction framework.

Typically, MDPs assume that Markov states are present, meaning that no further information such as history of previous states is needed to solve for and implement optimal decision-making policies. In other words, we want the state at a time step to be a good enough representation of the problem for selecting the best actions in each scenario (also known as the *Markov property*). Given the former assumptions are satisfied, an RL problem can be formulated as an MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. Note that in practice, however, the internal environment states may differ from the ones provided to the agent to improve convergence performance. Furthermore, while most RL algorithms imply Markov states, the environments and classes of problem tackled by this thesis are usually partially observable MDPs due to the high degree of uncertainty experienced[67]. Nonetheless, for simplicity, the Markov property is assumed to hold in most available RL implementations.

In episodic environments, these interactions continue until the agent reaches a user defined terminal state, at which point it restarts, although infinite horizon environments are also present in several applications (i.e., operating a building demand response system). In a flexibility planning application, an episode is equivalent to a simulation of the full project horizon. In training, the agent is fed a wide range of scenarios where they apply policy π through many sequential interactions, with typically randomized state initialization s^0. These interactions are at the core of the learning process, as over time the agent gradually improves the accuracy of expected value estimations for either each state $V_\pi(s^t)$ or state-action pair $Q_\pi(s^t, a^t)$ under policy π. Ultimately, the aim is to develop an optimal decision-making policy π^*, as captured in (2.1), which maximizes *discounted* accumulated rewards per episode. It is possible to develop either deterministic policies where the action is a direct function or the state ($a = \pi(s)$), or stochastic ones which yield a probability distribution over the actions available in that state ($\pi\ (a \mid s) = P\ [\ a^t = a \mid s^t = s\ ]$). This is further discussed in the following sections and application case studies.

$$\pi^* = \arg\max_\pi \mathbb{E}_{\tau\sim\pi(\tau)} \left[\sum_t \mathcal{R}(s^t, a^t)\right] = \arg\max_\pi \mathbb{E}_{\tau\sim\pi(\tau)} \sum_{k=0} \gamma^k R^{t+k} \quad (2.1)$$

The generic MDP formulation is very similar to a decision rule approach in flexibility analysis, since the *if-then-else* structure needs first to determine the current state (i.e., observation of uncertain variable like demand or price) and take on certain actions (e.g., expand capacity, shut down operations) that will lead to a future state, with the goal of maximizing a certain discounted reward (e.g., cumulative profits) over time. In their conventional form, however, MDPs are only well suited for low dimensionality action and state spaces, as Markovian transition matrices can scale exponentially, with significant degradations in computational feasibility of policy evaluation at each decision step [10,67,133]. The primary difference between MDPs in RL and DRL is highlighted in Figure 2.3, whereby the decision-making policy π is stored in a tabular format in the former, while it is approximated through a neural network with parameters θ in the latter - as further elaborated in 2.3.2. In MDPs, in fact, the output decision is obtained by looking it up in a table of predefined state action pairs, while in DRL the output decision is found by using the state observation as the input into a neural network parametrized policy, leading to more dynamic decision-making possibilities.

2.3.1.2 Model-based Value and Policy Iteration

Generally, "a policy π is defined to be better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states"[67]. The optimal policy $\pi^*$is thus the one that is better or equal to all the alternatives, implying there may be more than one optimal policy in some problems, all denoted as π^*. Historically, the RL problem formulated as an MDP employs either value or policy iteration to compute optimal policies, similar as for DP[67]. Note, however, that these methods assume that we have a perfect and complete model of the environments MDP. Consequently, they are both guaranteed to eventually converge to the optimal policy. While these were not directly implemented in this thesis, they are briefly described here as their logic and mathematical foundation is important to understand more advanced DRL agents and their policy optimisation process.

Policy iteration begins with the initialization of an arbitrary policy π. The first step then involves calculating performance of this initial policy π(s) through the state value function shown in (2.2), also known as the Bellman expectation equation. Secondly, we find an improved policy π'(s) by selecting actions that maximize improvements in the value function $V_\pi(s)$ using one-step look-ahead. This one step look-ahead approach is used to quickly estimate the performance of different policy variations aiming to improve the performance of subsequent iterations. The original policy π(s) is thus replaced by the improved policy π'(s), and the policy evaluation and improvement steps are iteratively repeated until convergence to the optimal policy π^*. This also guarantees convergence to the optimal state-value function, $V^*(s)$, for all $s \in S$, shown in (2.3).

In value iteration, on the other hand, the state-value function from (2.2) is iteratively computed. A random $V(s)$ is initiated and then updated at each step using one step-look ahead to seek maximum values. The primary difference to policy iteration is then that the maximum is taken over *all* potential actions in that state. Therefore, it combines the evaluation and improvement steps into a single update step, iteratively estimating all possible cumulative rewards. The final optimal value function from (2.3) is then the same as that obtained through policy iteration. This implies that they will also converge to the same optimal policy π^*, which follows the trajectory of the highest value estimates.

$$V_\pi(s) = \mathbb{E}_\pi[R^{t+1} + \gamma V_\pi(s^{t+1}) \mid s = s^t \,] \tag{2.2}$$

$$V^*(s) = V_{\pi^*}(s) = \max_\pi V_\pi(s) \tag{2.3}$$

2.3.1.3 Tabular Methods for Model-free RL

The state-action value function (or Q function) expands on the former by decomposing the value of each action in each state, as shown in (2.4). Assuming the existence of a *deterministic* optimal policy, the optimal state-action value function can then be found according to (2.5) for all $s \in S$ and $a \in A$. Crucially, the optimal policy and Q^* can be found in lack of an explicit environment model, differing from V^* which requires precise one-step look-ahead to reach the same objective. The state transition probability matrix, therefore, does not need to be know *a priori* but rather can be estimated or approximated through agent-environment interactions. This significantly expands the scope of problems which can be tackled with RL compared to model-based approaches.

$$Q_\pi(s,a) = \mathbb{E}_\pi[R^{t+1} + \gamma Q_\pi(s^{t+1}, a^{t+1}) \mid s = s^t, a = a^t \,] \tag{2.4}$$

$$Q^*(s,a) = \mathbb{E}_{\pi^*}\left[R^{t+1} + \gamma \max_{a^{t+1}}(s^{t+1}, a^{t+1}) \,\middle|\, s = s^t, a = a^t \,\right] \tag{2.5}$$

Consequently, temporal difference (TD) learning arises as an alternative to find the optimal policy by iteratively updating the estimates for value functions V_π and Q_π under policy π – essentially introducing learning from experience [67]. In the lack of an explicit environment model, it is able to do so by using *Monte-Carlo* methods for sample-based value and return estimation. Further, to address the large variance when estimating the distribution of returns using a single sample, bootstrapping methods tend to be implemented to approximate *n-step-returns* using the recursive nature of the Bellman equations discussed above. Updates are generally conducted on every action or timestep rather than at the end of the episode, allowing better differentiation between states and temporal correlations within the episode. The way these updates are conducted leads to a further differentiation in RL methods. Firstly, they can be made *on-policy,* whereby sample-based estimates are generated

according to agent-environment interactions under the specific policy that is being evaluated or iterated. Secondly, *off-policy* methods exist, whereby, generally both a behaviour-policy (for generating data) and a target policy (for actual reward maximization) are implemented for learning. One-step TD-learning and Sarsa are some of the earlier methods belonging to the latter, primarily used in policy evaluation by improving estimates of the value function[67].

Q-learning, on the other hand, broadly performs off-policy updates and is widely regarded as one of the most important breakthroughs in RL[67,135,136]. In its tabular form, it iteratively improves state-action value functions towards the optimal one by making the update shown in (2.6) and allowing the use of transitions sampled from any behaviour policy or past experience to improve estimates of Q^*. Note that α represents a positive *learning rate* and is a very important parameter is the design of even highly advanced RL agents, representing a trade-off between fast convergence, stability, and computational expense. In practice, however, small learning rates are typically implemented. Nevertheless, assuming all state-action pairs can be feasibly and continuously visited/updated, Q has been proven in former studies to always converge to Q^* [136,137]. The requirement to visit all state-action pairs introduces the notion of *exploration vs exploitation* through the behaviour policy implemented during training, as further discussed in 4.3.2.

$$Q(s^t, a^t) \leftarrow Q(s^t, a^t) + \alpha \left(R^{t+1} + \gamma \max_{a^{t+1}} (s^{t+1}, a^{t+1}) - Q(s^t, a^t) \right) \tag{2.6}$$

2.3.2 DRL Agents

2.3.2.1 Introduction

The applicability of methods discussed in the earlier section to real-world engineering design problems is limited by the practicality and memory requirements of learning a value for each state or state-action pair in highly dimensional and complex environments. Deep learning offers a promising alternative and is increasingly used to parametrize the RL problem. This has allowed the development of innovative algorithms to bridge the gap between these fields accelerate progress, enabling the efficient learning of optimal policies in more complex and uncertain environments. The state of the environment, also known as the *observation space*, is thus represented using a feature vector in DRL (as opposed to scalar values in MDPs and RL). This helps achieve better generalization through the combinatorial structure leveraged across features captured in the state space. It also introduces the importance of normalisation, particularly when features in the state space can vary significantly in magnitude. A parametrized NN approach (as further detailed in 2.2) can then be used to estimate either the value functions described in 2.3.1.2 and 2.3.1.3, or the policy itself. Accordingly, modern DRL algorithms are characterized as value-based, policy-based, or as actor-critic. It should also be noted that to improve training stability and reduce variance, many modern DRL algorithms utilise an *advantage function* in addition to the value functions described in the former section. This is generally

found according to (2.7) below, by subtracting a constant baseline (i.e. V_{π_θ}) for improved training stability – and providing some intuition on the relative value of each action in any given state under policy π_θ.

$$A_{\pi_\theta}(s^t, a^t) = Q_{\pi_\theta}(s^t, a^t) - V_{\pi_\theta}(s^t) \tag{2.7}$$

2.3.2.2 Value-based

This core of value-based DRL algorithms was first introduced by the seminal Deep Q-Networks (DQN) study[138], where the Q-values from (2.6) are iteratively approximated by a neural network with parameters θ. This greatly improved generalization ability as compared to tabular approaches. During the training process, an approximate action-value function is found by minimising the squared TD error based on each interaction and transition experienced. Note that this is different from the TD error calculated in (2.6), motivated by instabilities during the training process found in previous works using value function approximation[67]. These are driven by the sequential and correlated nature of interactions, which can lead to excessive variance across updates. Additionally, even very small updates to the Q-function can lead to substantial changes in the behaviour policy and resulting data fed into the learning process. Finally, changing action values are correlated to target values, which can lead to non-stationarity in the updates.

These issues were addressed by the original DQN work through two crucial modifications: experience replay and a target network. The former implements a memory buffer, where the results from previous interactions are stored and uniformly randomly sampled during training to update the parametric estimation of the action-value function Q_θ. This leads to reduced correlation between consecutive updates and the sequence in which states are visited. The latter, on the other hand, utilises a target network with parameters θ^- (taken from the online network parameters θ several steps in the past) for bootstrapping updates. This reduces the correlation between action selection and value estimation steps, through the introduction of a delay between them. These changes allowed DQN to achieve above-human level performance in the complex Atari domain, learning to play games purely from interactions with pixel inputs[138].

Several other extensions to value-based methods have been developed since the original DQN agent. Double Q-learning was implemented to further address overestimation bias of Q-values and decouple it from action selection by utilising one network to select the next greedy action and another to evaluate it[139]. A dueling network architecture was introduced to improve the efficiency of value estimations, separating approximation of the state-value function and advantage function [140]. This also allows for improved differentiation of the importance of each action in each state, particularly in the presence of similarly valued actions, as well as enhanced generalization in highly dimensional action spaces. Prioritized experience replay leads to more efficient training, by placing a stronger emphasis

on transitions which generate a greater learning error, as opposed to the uniform sampling by DQN[141]. Distributional value functions have been proposed, whereby the expectation of return is estimated by a probability distribution rather than a scalar value[142]. This can be particularly advantageous in highly stochastic environments, where performance in the less likely tail ends of scenarios can be significantly skewed from the average one. The Rainbow agent combined all the advancements above to yield state-of-the-art results on the Atari domain, although previously criticized as difficult to implement in practice[143]. Bayesian DRL may also be used to incorporate prior knowledge of the problem, when available, and improve uncertainty characterization[144].

2.3.2.3 Policy-based

The other approach to solving the DRL problem requires directly parametrizing the policy π_θ via an approximated NN function with parameters θ. As such, this class of algorithms does not need explicit value estimates for selecting an action based on a state. Generally, these methods exploit gradient information to maximize expected return $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T} \gamma^t r_t]$ by updating policy parameters θ, as shown in (2.8). Here, $\nabla J(\pi_{\theta^i})$ shows the stochastic approximation of the policy gradient based on its parameters and the selected performance measure. An analytical expression for $\nabla J(\pi_{\theta^i})$ was first introduced via the *policy gradient theorem*[145], although in practical problems targeted by this thesis it is infeasible to evaluate it directly, thus it is iteratively estimated through agent-environment interactions under policy π_θ. Starting policy network parameters θ_0 are normally initialized randomly, and are updated over the training process to eventually yield the optimal ones θ_*. Figure 2.4 (**i**)[146] shows what may be an example non-convex policy function which needs to be optimized and is captured by a simple two-layer NN with parameters θ^0 and θ^1 in Figure 2.4 (**ii**), respectively. The example policy search is initialised at point A while the optimal one located at point B in Figure 2.4 (**i**). Figure 2.4 (**iii**) then generally captures the DRL policy optimisation process to go from $\boldsymbol{\theta_0}$ to $\boldsymbol{\theta_*}$ by learning over a large number of agent environment interactions.

$$\theta_{i+1} \leftarrow \theta_i + \alpha \nabla J(\pi_{\theta^i}) \tag{2.8}$$

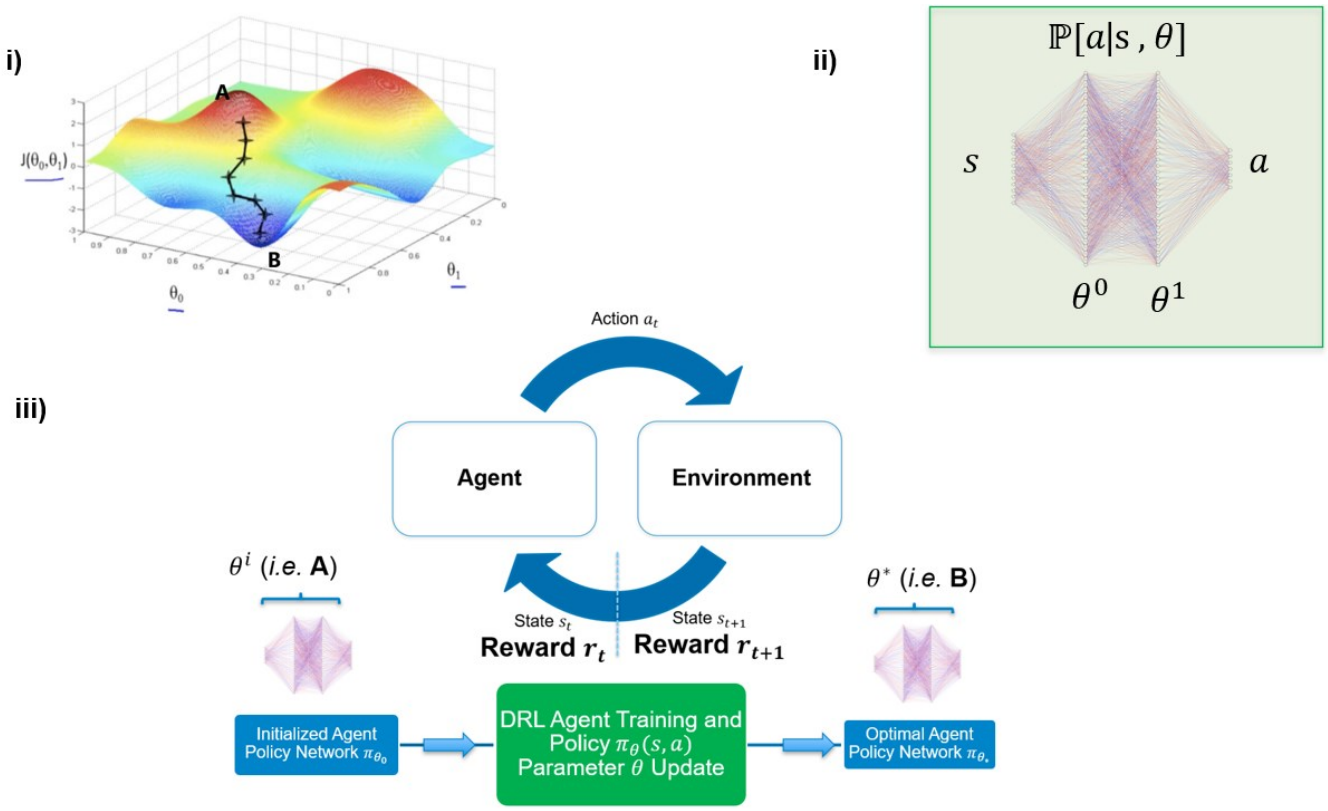


Figure 2.4: Graphical overview of policy gradient-based DRL optimisation process. Part **i)** shows an example policy function which we would seek to optimise, highlighted by the presence of many local minima between the initialization point A and the global optimum B. The parameters for which we seek optimal values in this case are the neural network weights θ^0 and θ^1 shown in Part **ii)** of the figure. Finally, Part **iii)** gives an overview of the training and optimisation process of a policy neural network when randomly initialized to point A in Part **i)** and seeking to reach point B in the same.

The foundational policy gradient algorithm is REINFORCE, employing empirical return estimates for a trajectory under policy π_θ[147]. In order to reduce the variance associated with this, several novel methods have been proposed in recent years. Guided policy search (GPS), for instance, accelerates learning by providing the agent with good example sample trajectories (i.e. sequence of actions made in states) [148]. It effectively biases the policy-optimisation process by optimising policies to match the sample trajectories more closely, while still maximizing returns. The need for sample trajectories, however, has limited its applicability to domains where it is feasible to know beforehand which actions in which states can provide a strong performance baseline. Trust regions are a more popular approach, constraining potential policy updates during optimisation to not be too far from the previous one. This reduces the chances of making policy updates which the agent cannot reasonably escape from, and can guarantee monotonic performance improvements[69]. Several metrics can be used to estimate the difference between two policies or the trust region magnitude, although original works computed the Kullback-Leibler (KL) divergence[67].

More recently, Trust Region Policy Optimisation (TRPO) was proposed to better handle higher dimensional environment inputs in continuous domains[149]. The optimisation procedure uses importance sampling of the weighted advantage estimate and constrains a quadratic approximation of the KL measure. Generalised advantage estimation is an extension of this aimed at reducing training variance[150]. TRPO, however, can be very computationally expensive as it requires the approximation of second-order derivatives. Proximal policy optimisation (PPO) is the most commonly implemented alternative, as a simple adaptive KL penalty or heuristics are implemented to modify the objective function, leading to an unconstrained optimisation problem based on first derivatives only[151]. As such, it is generally less expensive than TRPO at achieving similar performance, with more limited need for hyperparameter calibration across domains[152].

Many more agents exist based on policy search, and their learning is typically iteratively estimated by combining TD-learning concepts with backpropagation and SGD-like methods. Gradient-free approaches such as evolutionary strategies are also possible, which are better suited to handle variance in action frequency, delayed rewards, and long-horizon environments- although typically only computationally feasible for the optimization of smaller NNs with limited number of hidden units [153]. Nonetheless, they have been shown to be less sensitive to highly disconnected search spaces with non-differentiable function approximators. As such, they present a promising alternative to avoid gradient computations in some environments, even if lacking convergence guarantees. They work through the random generation and recombination of policy network candidates based on their fitness function, as shown in Figure 2.5 (**i**) for the same example two-layer NN depicted in Figure 2.4 (**ii**). Figure 2.5 (**ii**) depicts how candidate policies are first randomly generated with this approach (top) and how they then explore a greater region of the search space during the policy optimisation process through stochastic search like-methods (bottom) – although any other "black-box" optimisation approach may be implemented for this purpose.

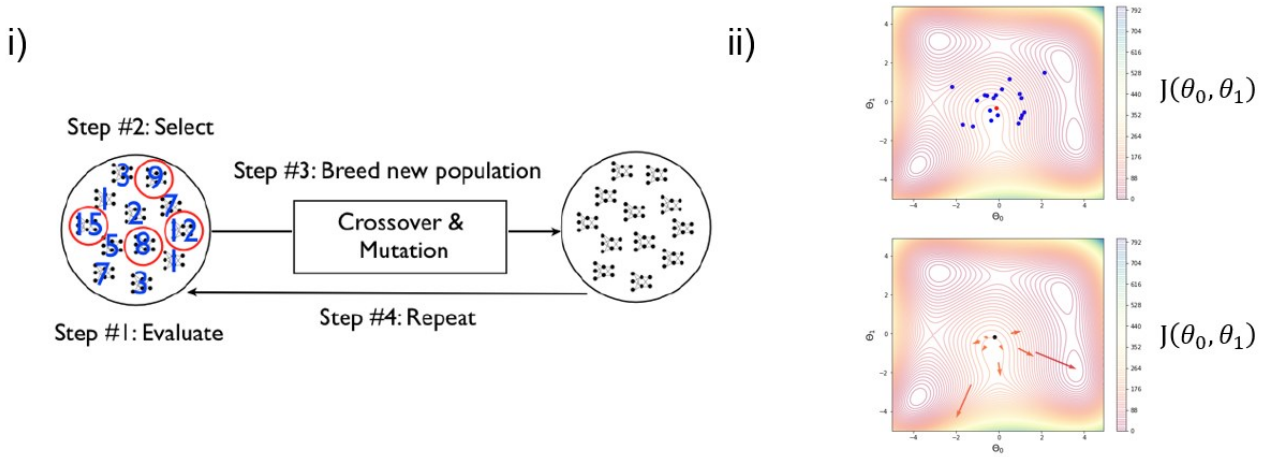


Figure 2.5: Graphical overview of gradient-free approaches in policy-based DRL

2.3.2.4 Actor-Critic

Aiming to combine some of the benefits from both value and policy-based methods, actor-critic approaches normally maintain an explicit value-function NN approximation for baseline computations, although they still use this to compute policy gradients. A significant benefit of this is the ability to use off-policy data on-policy gradient updates, and the general logic behind how they work is captured by Figure 2.6. Advantage Actor Critic (A2C) is one example, implementing asynchronous methods and parallel environments to update a stochastic policy. Deep deterministic policy gradient (DDPG), on the other hand, presents a deterministic actor and policy optimisation process[154]. More recently, Soft Actor Critic (SAC) is an off-policy method with on-policy corrections, employing entropy-based reward augmentation for exploration and stochastic value gradients[155]. While there are several other methods belonging to this class which cannot be discussed here due to length limitations, please refer to [156] for a more detailed overview of actor-critic methods and the implications for sample complexity when using function approximation.

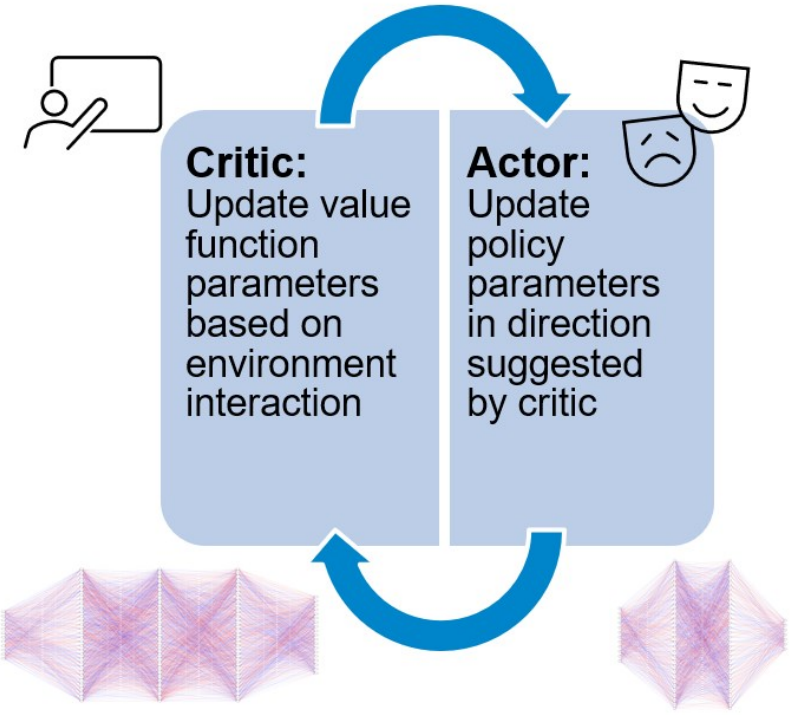


Figure 2.6: Overview of actor-critic method for DRL policy optimisation.

2.3.2.5 Summary

The review on DRL algorithmic literature suggests there is no one size fits all solution, and different agents have been found to perform better across various domains, with the field continually evolving.

While value-based methods have achieved significant success in several applications, challenges remain. High-dimensional state spaces and the trade-off between exploration and exploitation pose

ongoing issues. Additionally, improvements are sought to enhance sample efficiency, stability, and the ability to handle continuous action spaces. Integrating value-based methods with other techniques, such as model-based approaches, meta-learning, or hierarchical reinforcement learning, will be necessary for continued advancement of the field.

Policy gradient or actor-critic methods overcome several of the limitations of value-based approaches, but the best performing algorithm remains problem and domain specific. Specifically, they tend to present better convergence properties in high-dimensional problems with a long horizon. They are also compatible with continuous domains without the need for modifications, expanding the scope of DRL. They can produce stochastic policies, which yield a probability distribution over actions in a state. As such, they much better suited for decision-making under uncertainty than the implicit deterministic policies obtained from value-based methods, particularly in a flexibility planning application. On the other hand, they tend to be less sample efficient, can present higher variance during the training process, as well as are more likely to be stuck in local optima due to the more challenging policy function which is optimised. Gradient-free approaches such as evolutionary methods present another promising alternative, but are severely limited by their computational expense when optimising over a large number of parameters in deeper NNs.

The specific agents used in different chapters of this thesis and the motivation are further discussed in more details in their respective sections.

2.3.3 Practical Implementation

DRL agents and networks can be implemented through several frameworks and packages, depending on the level of customization desired by the user. For more advanced DRL practitioners, general deep-learning framework level abstractions can be used to build agents, environments, and neural network function approximators with more flexibility on implementation details. The primary open-source frameworks which have driven recent advances in the field are Keras[157], TensorFlow[158], PyTorch[159], and JAX[160]- among many others. Generally, TensorFlow is the most popular choice for industry implementations while PyTorch is preferred in research settings as it allows for easier experimentation and customization.

DRL applications to more practical case studies (i.e. outside of testing performance in the Atari domain) have also been heavily based on easy to implement libraries, where agents are normally already coded. These are mostly compatible the OpenAI Gym framework, the standard way to design and code environments in Python[161]. Once a suitable environment is developed, several different agents, network architectures, and hyperparameter configurations can be quickly experimented with and compared using DRL libraries. The most common include rllab[162], OpenAI Baselines[163], TensorFlow Agents[164], Dopamine[165], and Stable Baselines[166].

Even considering the abundance of open-source DRL packages, hyperparameter tuning remains a crucial component in agent training and final performance achieved. Several open-source packages are available for this purpose, including Ray-Tune[167], Hyperopt[168], Talos[169], and Optuna[170].

The specific implementation details used in this thesis will be further discussed for their respective case studies, but are primarily based on the libraries and frameworks presented in this section.

2.3.4 DRL for Engineering Design

In order to maximize the impact DRL has in the field of engineering design, it is essential to understand the kinds of problems it can handle and where it marks an improvement from current methods. This may be the reason why, while its potential is widely recognized across academia, it has exhibited limited implementation on real-world engineering problems [171]. Even considering recent breakthroughs, the use of a fully automated agent for the design of a complex and expensive infrastructure project remains far-fetched. The majority of DRL implementations thus far in the domain of engineering have centred on more classical control problems which can be more easily fit within an MDP and stimulate agent learning[69]. In particular, several advanced implementations have been recently presented aiming at optimal process control of bio-reactors and other chemical engineering applications, with the authors showing very impressive results[172-174]. A recent review of DRL in energy systems management suggests increasing potential, although primarily focused on control to maintain appropriate technical parameters during operation[175]. Engineering design specific applications have, therefore, normally focused on a specific subset of the problem, as in classical ML.

A general framework for design optimization using DRL was proposed in [176] and was applied to the air foil angle of attack for a hypothetical stator with varying rotor designs. The work in [177] also shows the strong performance of DRL in the design task of fatigue resistant solutions for ship design compared to evolutionary methods. Both implementations, however, maintain limited scope within the overall problem, with overly simplistic reward functions or action spaces. There are a number of works focusing on new RL algorithms or techniques to allow combination with more complex problems, however they tend to focus on the mathematical or algorithmic side[67,83,178,179]. More promising is the optimization performed in [87], where off the shelf DRL algorithms are implemented to design a microfluidic device for flow sculpting. Using a Double Q-Network architecture combined with Hindsight Experience Replay, the DRL agent produces designs that are consistently in the 99th percentile of the entire design space, compared to those yielded via a much more computationally expensive brute force approach. The authors also show DRL can be effectively combined with human insights via a targeted refinement step in order to manually guide training (i.e. GPS) when desired and accelerate convergence and optimization thereafter.

The potential for DRL to the aim of this project was perhaps most thoroughly investigated by [86], where it was first implemented to simultaneously solve the competing energy systems basic design

and dispatch strategy. In their approach, the dispatch problem is presented as a sparse reward problem, compared to the traditional formulation as a distributed reward problem in a number of previous energy management applications[127,180-182]. The researchers used both a fully connected and CNN architecture, whereby the convolutions were introduced to allow consideration of future forecasts of operating conditions. The DRL solutions, particularly for the CNN implementation, are found to outperform white box or grey box models by up to 60% in some cases, with significantly lower computational time. This enhanced performance becomes clearer as a higher degree of complexity in energy flows is introduced to the design problem. The best performing solutions are able to minimize the use of dispatchable sources and create a more autonomous energy system from the grid. However, the configurations produced ultimately show there is still a significant room for improvement in this methodology. Most of the solutions yielded a higher amount of dispatchable generation than annual demand (due to load factor efficiency economies), with over 20% of renewable generation not utilized in most configurations.

Better integration of the design phase into the dispatch solution evaluation is suggested in the work as holding significant potential for improved performance. Additionally, there was no consideration of decision-making flexibility through the project duration to adjust capacities over time based on realized uncertainties. In fact, to these authors' knowledge, there has been no previous study exploiting the potential of DRL in the context of enabling flexibility in engineering systems design.

Chapter 3

3 Methodological Definition: Data-driven Flexibility in Design

This chapter defines and presents the novel methodological approach developed in this thesis by aiming to address the following objectives from 1.6.

- RQ1: How can ML and other data-driven methods add value across the different phases of the Flexibility in Design Framework?
- RQ2: Can DRL be used to analyse flexibility in engineering systems design under uncertainty as part of a standardized process?

It begins by generally exploring how ML and data-driven approaches could be used to augment a Flexibility in Design framework. It follows by presenting the innovative DRL based flexibility analysis procedure. An overview of the problem formulation process is then presented, including practical considerations such as how to identify applicable systems where this may be particularly helpful. Finally, a holistic summary of the overall methodological procedure is presented for implementation on real-world engineering systems, based on the various components and factors explored. The work presented in this chapter is primarily based on the following peer-reviewed publications[107,183] outlined in full in 1.8. This novel methodology and framework are then applied and tested on several very different case studies in the following chapters.

3.1 Data-driven Flexibility in Design Framework

3.1.1 Overview

The argument we make in this chapter is not for ML to be used for the standalone design of flexible engineering systems or to generate mathematically more optimal solutions, but rather as an important tool that can help practitioners complement standard methods in order to achieve better performing solutions. The seminal framework presented in 2.1.2 is used as a baseline to explore how ML and other data-driven approaches may be used to enhance performance across the phases involved. The review from Chapter 2 suggests that there is potential for the development of a novel tool for flexible design supported by advances in ML techniques. Current methods (detailed in 2.1) available to designers and engineers are well suited for the enhancement of system flexibility compared to a traditional or baseline design process, but present a number of analytical and practical limitations. They tend to rely on either domain expertise or generic approaches and thus are limited in the number

and type of solutions they are able to generate or consider based on what has previously been done or is known to decision makers. There is likely solutions and configurations embedded within the increasingly available data, from sources such patents and publications, which can increase delivered expected value by a project that are not being considered by engineers and decision makers.

Several potential avenues were thus explored and tested, although it is recognized that there may be additional roles for data-driven approaches outside of the ones mentioned here. Nonetheless, the most impactful integrations are deemed to be during the uncertainty recognition, concept generation, and design space exploration phases. Figure 3.1 addresses this in greater detail, showing the potential integration of ML in each of the relevant phases, as well as the corresponding impact or output[107]. An intuitive motivating example is presented in 3.1.2, followed by a more detailed discussion of the role and impact of data-driven approaches in each of the phases depicted in Figure 3.1.

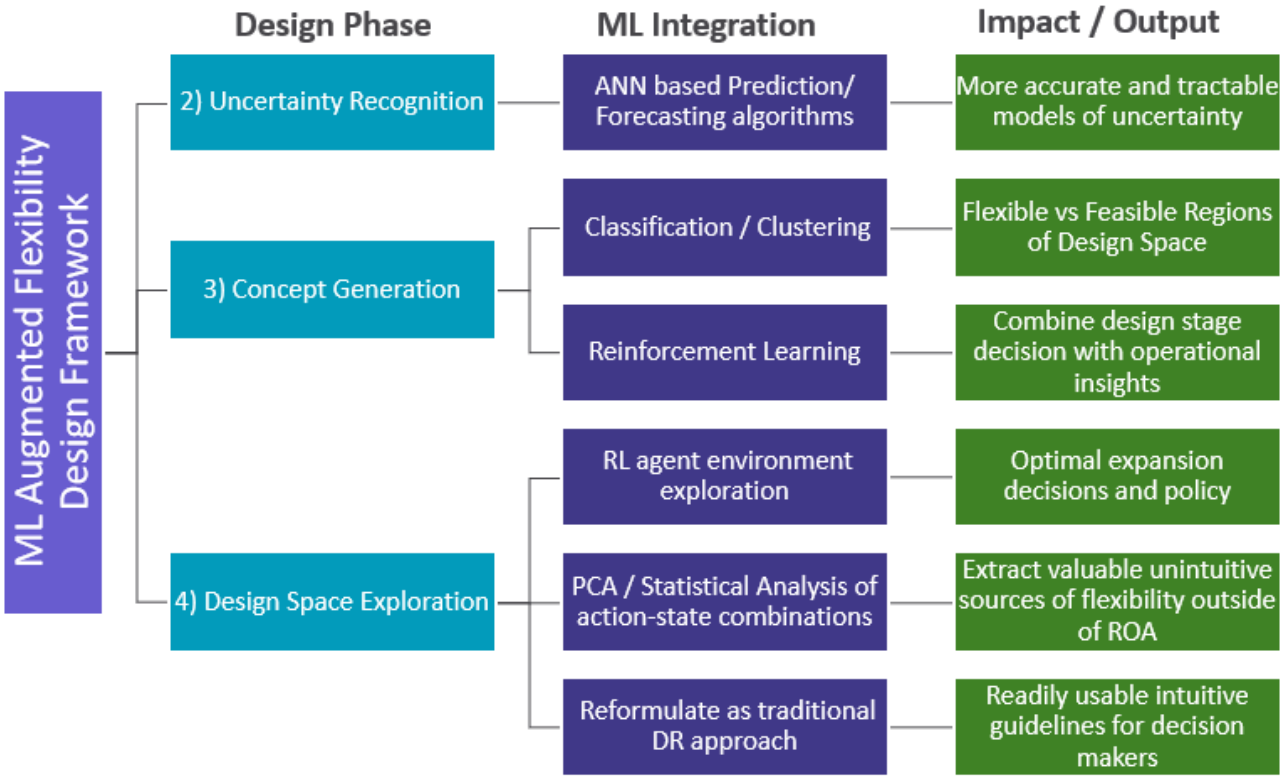


Figure 3.1: Data-driven Flexibility in Design Framework

3.1.2 Intuitive Example Application

The motivation and potential of the proposed framework is illustrated through a simple case study of the design of a multi-story parking garage subject to demand and price uncertainty, first presented in [184], where developers need to decide how many parking spaces to build. This is a foundational paper to explore the relevant concepts in the field of Flexibility in Design, which has since been

benchmarked against many more advanced approaches. Note that this case study is also used for methodological validation in Chapter 4, where it is explored in greater detail.

The inflexible baseline would be to construct everything at project start, while the flexible option starts with a smaller capacity but with expansion possibilities, albeit at a higher cost. There are a number of possible alternatives which could be considered that ultimately deliver the same flexible expansion capability, such as inclusion of larger foundations, stronger material, or fully modular implementation (i.e., stacking prefabricated levels on top of each other). No specific ML integration is included for phase 1 (Baseline Design), however any ML method applicable to general engineering design could be used during this stage to define the optimal inflexible baseline. During phase 2, ML could be used to build more accurate models to understand how price and demand uncertainty will evolve over time and help make better informed design decisions on how flexibility could be most impactful. This could lead designers to try to focus on either generic capacity expansion, or for example target expansion for pricing past a certain threshold only. Surrogate modelling may be implemented at this stage to determine expansion cost function from a dataset of previous expansion decisions in other projects more accurately and easily than via analytical means.

ML classification or clustering algorithms could then be used to identify the most flexible regions of the design space, rather than simply the feasible ones, and understand the relevant trade-offs during concept generation. The clustering criteria could be varied for a more holistic evaluation, in order to assess for example regions of the design space that are well suited to respond to price but not demand uncertainty, and vice versa - and the resulting design implications. Similarly, this could be extended to efficiently identify regions of the design space better suited for sustainability or resilience objectives. Based on the results from phase 2, decision makers could choose to focus on designs allowing adaptability to demand uncertainty, for example, and look at their common value adding features.

In terms of RL applications, there is immediate potential for supporting the decision rule method from flexible design, particularly during phase 4 as shown in Figure 3.1. A traditional decision rule approach for the multi-story garage design could be formulated as (IF demand for last 2 years exceeds current capacity, THEN expand by 1 floor, up to max allowable capacity). Once flexible design concepts and strategies are generated in phase 3 and all relevant actions recognized within the RL environment, the training of the agent would be launched. In the garage case, the starting action would include the decision on starting levels of floors and flexibility enabler (i.e., stronger columns or prefabricated containers). This would help integrate the early design decision with operational insights for responding to uncertainties. It could also be better suited to produce combinations of decision rules by more systematically exploring the design space for value enhancing measures.

For example, this RL based exploration may suggest building 1 floor only once capacity reaches a certain level to avoid investment risk, while building 2 or 3 floors up to that level in order to ramp up

demand more quickly. Another purely theoretical example in the garage case may be that the agent finds building an extra half floor on some expansion years as part of its optimal policy, which would likely not have been considered in a domain expert defined flexibility analysis. This could allow for the garage to ramp up capacity more quickly if demand greatly exceeds expectations on some years, reducing lost revenue during the building period and reducing liquidity risks. Advanced statistical analysis of optimal action-state combinations could help yield a more "explainable" process. These value enhancing actions could then be re-formulated as a traditional DR approach to provide some readily usable guidelines to decision makers.

3.1.3 Uncertainty Recognition

Increasing accuracy and understanding of relevant uncertainties is essential in any engineering design problem. We argue, however, that they are particularly important when designing for flexibility as the objective is to maximize adaptability to realised uncertainties. In the simplest of cases, ML methods may be used to generate more representative input scenarios and parameters to be used in scenario planning, decision trees, or other relevant techniques mentioned in 2.1.2. When accounting for highly complex physics-based problems, surrogate modelling may also be helpful to simulate scenarios, helping to understand potential random shocks from external factors at a greatly reduced computational expense, as shown by [185]. This may also allow a much higher number of scenarios to be evaluated within the design process. Historical data on trends for the uncertain variable would, however, likely be necessary for improved accuracy here, which may become a limiting factor in a lot of applications. Consequently, common issues such as bias or overfitting would need to be evaluated carefully when using data-driven methods for uncertainty recognition.

3.1.4 Concept Generation

Appropriate clustering and classification implementations during phase 3 could greatly accelerate the design process. They may be used to provide a list of common features for previous system designs well suited to deal with a particular uncertainty identified from phase 2, or to estimate how a specific design compares to other ones in the field. The advantage of ML here is that it may be able to do so in a much more computationally feasible manner than alternative techniques. Furthermore, it is reasonable to assume that flexible designs for a particular system are likely to share some common features (i.e., stronger columns for garage or oversized inverter for solar PV system) not found in their inflexible counterparts. These methods would allow for much faster identification of these features and for designers to choose to prioritise them if needed from very early stages. It may then be possible to build a predictive model that is able to generate a flexible design solution based on an inflexible baseline or a few basic inputs, for example only specifying max initial investment and local material costs, based on combinations of ones previously found within historical datasets of multi-story garage designs. The integration of all these applications, however, would necessarily require a large data set of system designs, both flexible and not. This is not normally available for most design problems;

however, it is something that could be developed over time if designs were made more accessible or even built artificially using techniques such as Generative Adversarial Networks (GANs) [186].

There are also a number of potentially interesting applications more closely related to the field of computational creativity. For instance, ML may be used to develop novel metrics measuring flexibility of a system design by looking at a large number of candidate designs. These novel metrics or evaluation functions could then be used by designers to initially estimate the flexibilities of different potential solutions to be considered for further evaluation. These ways of estimating flexibility may be unintuitive to practitioners in the field but could uncover potential value enhancing methods or configurations that do not appear flexible at first, or vice versa. For example, using a stronger material may seem advantageous at first but can lead to vulnerability to the future price of the material when expansion decisions are made, thus lowering the actual value of flexibility in the system compared to using a larger amount of a weaker material subject to less uncertainty.

During the early stages of framework development, the use of a modified Engineering Systems Multiple Domain Matrix (ES-MDM), was explored as a potential format for more holistic approach to concept generation. This sought to use more complex input data to capture multidisciplinary systemic dependencies in the design process, and their impact on Sustainability (S), Resilience (R) and Flexibility (F), presented at the Supergen Energy Networks Hub Risk Day[183]. The impact category was included merely as a qualitative indicator to guide the DRL agent in design space exploration, balancing competing objectives as shown in Figure 3.2. In fact, it is normally unreasonable to quantify the full spectrum of these interactions early on in a project's lifetime. Based on the high degree of complexity and data-requirements associated with the ES-MDM, it was not further pursued.

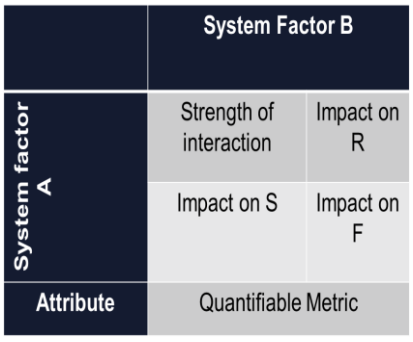


Figure 3.2: Modified ES-MSM. The impact on S, F, R represents a qualitative assessment by domain experts of the perceived effect of the interaction ranked as either positive, negative, or neutral.

3.1.5 Design Space Exploration

Given the requirements and objectives of Flexibility in Design, RL is proposed to likely be more impactful and effective than traditional ML. The sequential decision-making nature of the method is well suited to examining impact of design and subsequent operational decisions on the performance of

the system. As such, the impact of exercising different types of flexibility over time can be quantitatively assessed over a large number of scenarios. Sequential decision-making nature refers to the RL agent exploration of the environment, rather than the design process itself. It is thus expected that the design process will remain more iterative using the proposed framework, where insights gained from RL exploration over time can be collated to inform design decisions, rather than modelling the agent as the designer directly.

We make the case that RL holds great potential to solve combinatorial optimization problems such as the ones found in flexibility decision analysis, as the reward stochasticity found in decision making and the reduced reliance on human intuition inputs allows it to explore more solutions than possible with comparable methods. For example, it could be used in combination with decision rules to create more dynamic and adaptable systems under uncertainty. Looking at a standard implementation, decision rule variables and their potential feasible ranges are normally identified at start by domain expertise, and their optimal values determined through simulation. In an RL based approach, these would not need to be identified at start but rather just all possible actions recognized within the environment. Over a much larger number of scenarios than possible with standard methods, the RL agent could then develop their own decision rules to maximize utility in an environment which are not limited to the form chosen by the designer at start (i.e., IF-THEN).

Neural networks allow the development much more complex value function representation and enhance understanding on when and how to implement different and new types of flexibilities. Furthermore, a DRL implementation is better able to handle conflicting objectives and multiple sources of uncertainty than typical decision analysis methods, thus is more representative of the engineering design setting. This methodology could also enhance exploratory creativity, as the input values for design variables are much less restricted than in standard optimisation procedures, given the advantages of parallel computation and resulting algorithms. The exploration of the design space could also be easily extended outside the economic realm in order to enhance system sustainability and resilience by integration directly into the reward function. Furthermore, the theoretical common ground shared with standard methods in decision analysis allows for a more straightforward solution procedure to be created out of it, such as the decision rule one presented in 2.1.5.

3.2 Deep Reinforcement Learning Flexibility Analysis

3.2.1 Introduction and Background

The framework presented in 3.1 shows that, while there is great potential for integrating data-driven approaches in this field, using DRL as an alternative method during design space exploration holds the greatest potential at the time of writing. The objective is then to find the optimal combination of *initial design* configuration and *operational strategy* (policy) which allows engineering system adaptability to uncertainty through a learning process over a very large number of interactions with

the environment. In essence, the approach allows one to formulate the fundamental Flexibility in Design problem in a sequential-decision making manner to maximize (or minimize) a *user defined reward* signal. The reward may be based on several metrics, including net present value (NPV), CO_2 equivalent emissions, unmet load, total energy production, or problem-specific resilience measures.

The building blocks and primary decision-making components for DRL and MDPs, reviewed in 2.3.1, are graphically summarized in Figure 3.3. They provide some simplistic intuition on their function through the example application study from 3.1.2, further detailed in 4.1, by illustrating the different roles of the policy and value function in formulating flexibility strategies under uncertainty. The environment building blocks are thus used to create either a deterministic mapping of which action to take in which state (i.e. No Expansion or Expand 2 floors example) or a stochastic policy yielding a probability distribution over available actions. The economic value function may also be used in this context, by approximating the return of each action under different system conditions (for instance floors built and parking demand by number of people in Figure 3.3).

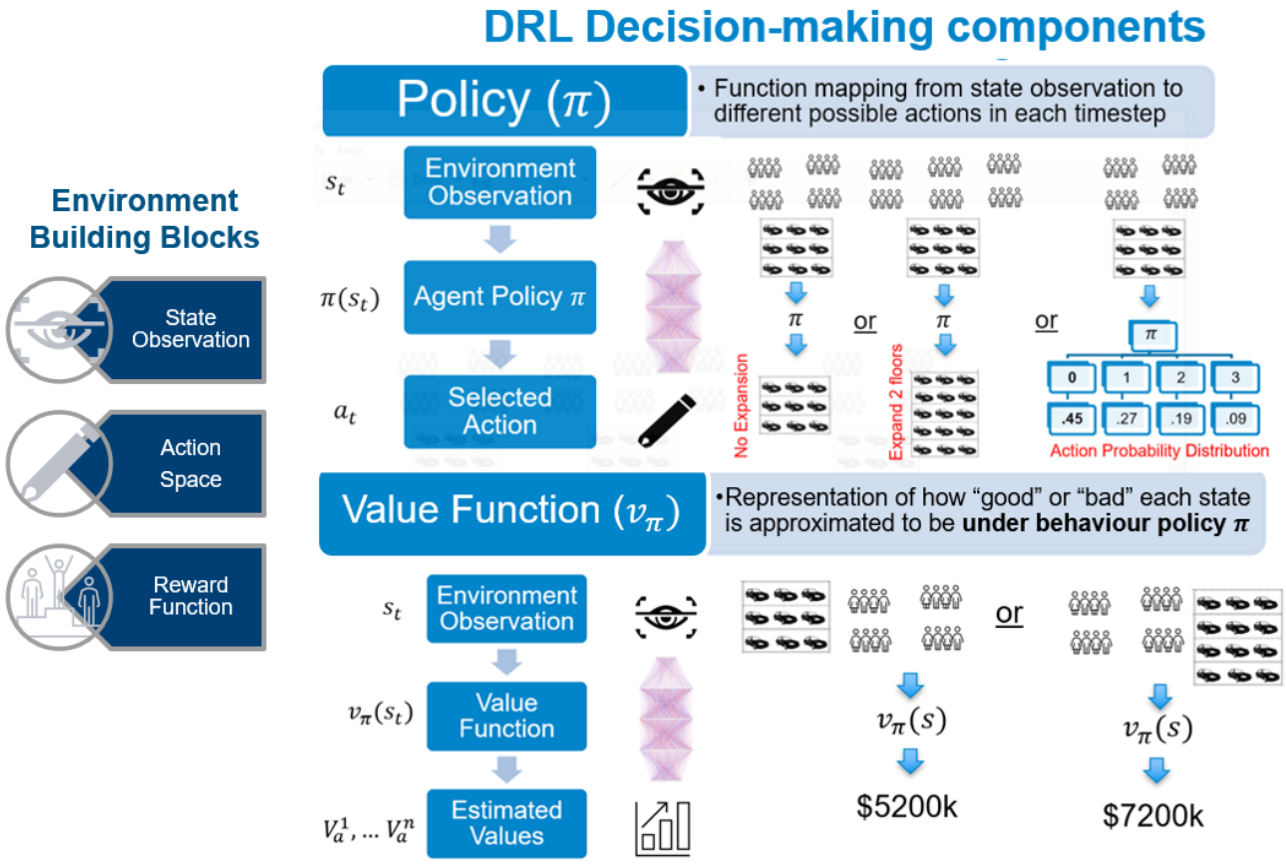


Figure 3.3: Overview of DRL decision making process on motivating case study. Note that action selection and value function estimates are simply for illustrative purposes and do not represent the actual optimised ones obtained through a complete DRL Flexibility Analysis.

3.2.2 Process Overview

The primary steps for the proposed approach are captured by Figure 3.4. In order to be most effective within the DRL framework, it is first essential to understand which variables best represent the system state at any point in time, as this is used to build the observation space for the DRL agent. This is not necessarily a straightforward exercise, as an excessive number of features (some of which may be redundant) can increase the complexity of the training process, while a more limited one may oversimplify the problem and not make use of all relevant information to develop flexibility strategies. Additionally, key performance metrics are identified in this phase, which will be at the core of the learning process depicted in Figure 2.3. The design of this reward signal is a critical component to the success of a DRL agents as it is its only true performance measure they are able to interact with for policy formulation and optimisation. The reward metric is likely to be based on economic signals for a lot of systems, although sustainability and resilience considerations can be directly integrated into the reward function to balance competing interest, depending on project specific objectives. Uncertainty quantification in phase 2 consists of two primary parts: variable identification and model generation. In variable identification we are looking for features which are deemed most impactful on the state of an engineering system over time and its resulting performance. Stochastic models are then built based on these key variables, and used to generate a large number of *i.i.d.* scenarios which characterize the DRL agent environment. In phase 3, the action space is created to deal with the uncertainties quantified in phase 2. It involves recognizing all flexibility enablers, starting from standard real options, as well as any other alternatives to reconfigure engineering system state during operation. Finally, we can systematically explore through the data-driven approach detailed in 2.3.2 what the optimal initial design and flexibility strategy is for the formulated environment.

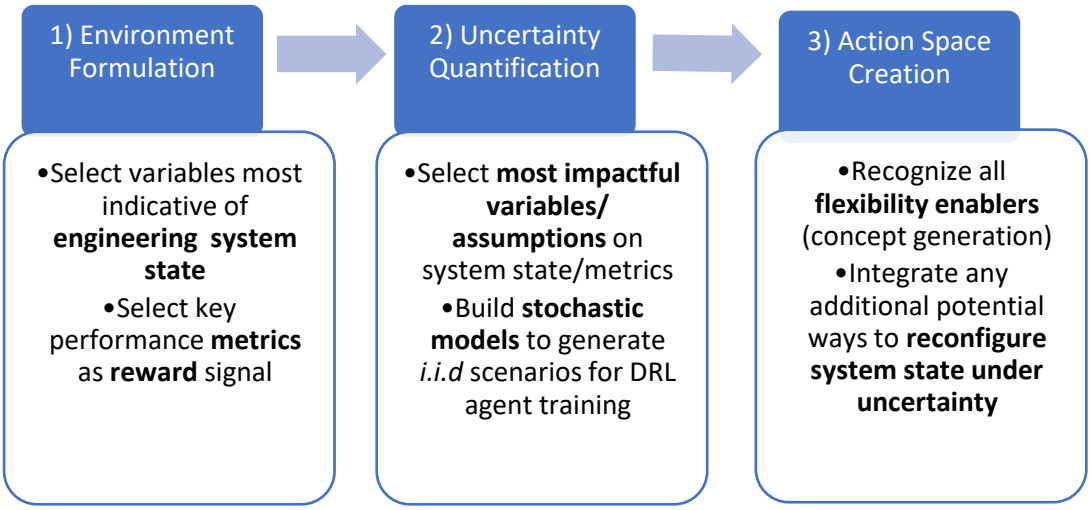


Figure 3.4: Primary steps in DRL flexibility analysis.

The major advantage as compared to alternative methods reviewed in 2.1 is driven by the power of NN function approximation of the flexibility strategy (or decision-making policy). Being less limited by the curse of dimensionality, it allows analysis of much more complex and realistic engineering systems models. As such, more sources of uncertainty and flexibility enablers can be tractably integrated into the optimisation procedure compared to alternative state-of-the-art methods. The potential to stimulate creativity on unintuitive solutions and complement existing methods is also created. Additionally, the impact of uncertainty on system state is estimated through interactions, meaning it does not need to be assumed before-hand. The end result is the development of a significantly more dynamic and generalizable decision-making policy under uncertainty, with improved potential for risk profile integration in early design stages. A high-level overview of how it may be used to exercise flexibility decisions in practice is further summarized by Figure 3.5.

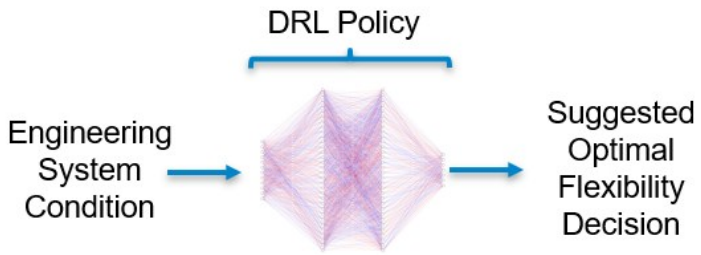


Figure 3.5: High-level summary of use of DRL policy in practice.

This DRL approach, never previously applied in this context, allows us to tackle the design problem in a manner more realistic to the actual conditions encountered. As a result, non-trivial engineering system evolution solutions may be discovered which present more optimal expected performance under a wide range of uncertainty realizations. The overall process proposed, broken down by technological and methodological selection, is given by Figure 3.6.

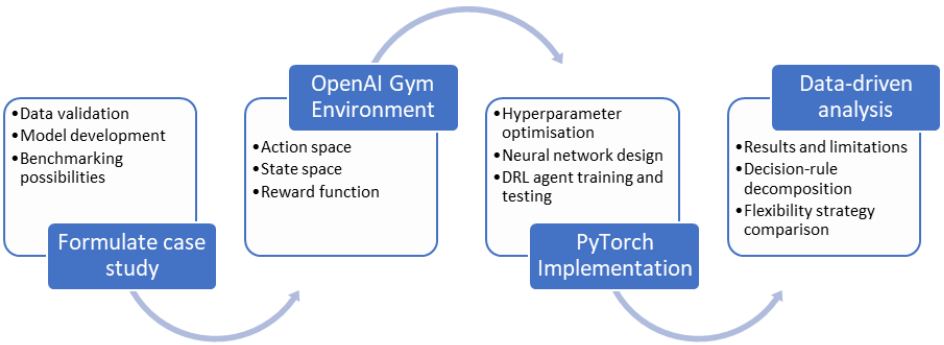


Figure 3.6: Overview of technological steps involved in DRL flexibility analysis.

It should be noted, however, that this is not suitable for all engineering design problems. In fact, the process presented in Figure 3.4 was also developed with the objective of determining the applicability of this approach on an example system. It stresses that it is focused on enhancing adaptability to uncertainty, but only in cases where it is possible to reconfigure system state during operation through some inclusions in the original design (i.e. flexibility enablers). Performing the exercise described by phases 1-3 can thus allow system operators or decision makers to identify if DRL flexibility analysis may enhance value in the case of their specific engineering system, or what changes would need to be made in order for it to become more compatible. Please refer to[171] for a more detailed review and discussion of the applicability of DRL on different system optimisation problems, including its shortcomings in certain domains.

3.3 Incremental Evaluation and Data Validation Procedure for Case Studies

In order to allow progressive benchmarking of validity of results, the incremental procedure depicted in Figure 3.7 is pursued for the case studies in this thesis. Firstly, a simplified and deterministic system model is created to represent most relevant variables and decision-making possibilities. The system definition is then expanded to resemble the real-life design problem more closely, with the introduction of its full range of stochasticity and all ways to reconfigure system state under uncertainty. Finally the performance of the produced design with the proposed approach is to be benchmarked against standard methods and an inflexible baseline where possible, including a full evaluation of various sustainability and resilience metrics relevant to the proposed solutions. It should be noted that while this incremental approach was implemented for all case studies, only results using stochastic conditions are presented in the following chapters, based on the objectives of this thesis.

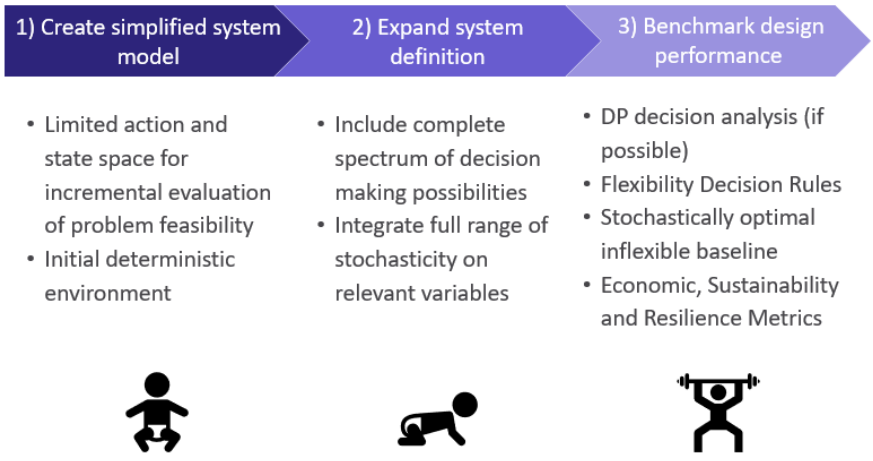


Figure 3.7: Incremental procedure implemented to evaluate the various case studies in this thesis.

The properties of the case studies presented in the following sections are captured by Figure 3.8. They show how the proposed approach enables for the analysis of flexibility in increasingly complex

systems. In fact, it can be seen that the dimensionality of both the action and state space is incrementally enhanced from Chapter 4 to Chapter 6, in line with the arguments presented earlier in this section. Similarly, the algorithmic approach grows in complexity based on the nature of the problems tackled, as will be discussed in more detail in their respective chapters.



Figure 3.8: Comparison showing incrementally increasing complexity of case studies evaluated. Photo credit for Chapter 4 is from[187], for Chapter 5 from[188], and for Chapter 6 from[189] .

It should also be noted that for all case studies presented above, results and data were rigorously evaluated to the ensure they were representative of the real world conditions. The case studies in Chapter 4 and Chapter 5 were taken from published journal papers, using all the same assumptions and formulation. For Chapter 4, the data was from the development of a real world infrastructure system in the UK, and the designs obtained through the proposed approach were extensively benchmarked with existing solutions. Similarly, the case study in Chapter 5 was commissioned by the NEA in Singapore, and thus all data used is representative of the system design problem faced by the government. Accordingly, the results obtained through the approach presented in this thesis are also in line with the ones currently in place. Please refer to their respective chapters for a more detailed overview. Finally, the most rigorous data validation process was naturally required for the case study presented in Chapter 6 and Chapter 7, given its novelty. Accordingly, several steps were taken to ensure the data and assumptions used to be as representative as possible. A few examples are listed

below, although please refer to the respective chapters for more details on data validation in the Mongolia case study, to the extent that it was possible in this context.

- Interactive workshops and data collection/validation by local Mongolian consulting group[190]
- Analysis of institutional reports focused on the link between poverty, energy, climate change, and migration in Mongolia[191-207].
- Benchmarking of energy system solutions against commercial tools like HOMER, systems installed through the REAP energy access program, or energy potential atlases.

3.4 Generic Engineering System Strategic Capacity Planning Models

This section briefly describes the generic capacity planning model that is relevant to the objectives of this thesis and is applicable to all the case studies analyzed in the following sections. While this can be formulated deterministically, we only present the stochastic counterpart here for brevity and as it is more closely aligned with the added value of this methodology. It considers the practicalities of economies of scale and the resulting implications for the performance of an engineering system design, as extensively researched in the literature[11,53,62,104,208,209]. Furthermore, although capacities of different components are the evaluation focus in all case studies presented, the proposed approach can be easily extended to other decision-making possibilities, as further explored in Chapter 6. Note that the maximization of ENPV presented in this section may be easily substituted by the minimization of expected net present costs (ENPC) or alternative objectives, depending on the problem at hand.

3.4.1 Inflexible Baseline

The inflexible baseline is defined here as any design alternative not presenting expansion or reconfiguration possibilities after project start. Considering a discrete finite time horizon planning period of T, with discount rate λ, let $\xi = (\xi^1, \xi^2, \dots \xi^T)$ be a scenario of uncertainty, where ξ^t is a vector capturing the uncertainty observed in period t, thus able to account for multiple uncertainty sources and relevant variables. S defines the set of all possible uncertainty scenarios s used for evaluation, assumed to be finite at S_s for simplicity, with equal probability of occurrence p^s. I denotes the set of all system components whose capacity is considered. The stochastic capacity planning model objective with no expansion flexibility can then be formulated according to (3.1). ${\theta_i}^t$ represents the installed capacity for various system components i, invariant with time t and bounded by their respective maximum allowable values $\theta_{i,max}$, as shown in (3.2). $\mathcal{R}^t\left({\theta_i}^t, \xi_s^t\right)$ is the revenue function and $\mathcal{C}^t\left({\theta_i}^t, \xi_s^t\right)$ the cost function for the system capacities considered in period t, which are based on both rigid system configurations and uncertainty realizations. Initial investment costs (aka *CAPEX*)

for each of these components ${C_i}^0$ are calculated through a power cost function accounting for economies of scale (EoS), with K a constant coefficient parameter, and α the EoS factor, as shown in (3.3). They are included as a zero-stage decision in (3.1). The magnitude of the EoS effect (as measured through K and α) can be estimated through statistical analysis of historical data on systems costs at different capacity levels. Alternatively, domain expertise combined with trends and cost data for similar engineering systems can be used[10].

$$Max\ ENPV = \frac{1}{S_s}\sum_{s=1}^{S}(-{\mathcal{C}_i}^0(\theta_i) + \sum_{t=1}^{T}(\frac{1}{1+\lambda})^t[\mathcal{R}^t({\theta_i}^t, \xi_s^t) - \mathcal{C}^t({\theta_i}^t, \xi_s^t)\ ]) \tag{3.1}$$

$$0 \le {\theta_i}^t \le \theta_{i,max} \tag{3.2}$$

$$\mathcal{C}^0(\theta) = K(\theta)^\alpha \tag{3.3}$$

$$\xi_s^t \ge 0\ \forall\ t \in T, s \in S, i \in I \tag{3.4}$$

3.4.2 Capacity Expansion with Flexibility Decision Rules

Decision rules can be embedded in the system model to act as triggering signals for exercising a particular flexibility in operations[11,62,80,81]. Based on the information revealed in each time step, decision rules enable decision-makers to sequentially decide when to proceed with capacity expansion, one of the primary strategies for enabling better adaptability. The decision-maker must decide on capacity deployment as part of a set of feasible decisions X_t for period $t \ge 0$, including a zero-stage decision on initial capacity, before any uncertainty is revealed. Let $X \subseteq X_1 \times \cdots \times X_t$ denote the set of all feasible capacity decision sequences $\boldsymbol{x}$, where $\boldsymbol{x} = (\theta^0, \theta^1, .., \theta^t)$, and let $\xi^{[t]}$ represent the full history of the uncertain variable up to time t. A decision rule, also known as an implementable policy, $\mathcal{F}$, is a function that maps each scenario of uncertainty ξ into a sequence of decisions x in X (i.e., $\mathcal{F}: S \to X$). Accounting for non-anticipativity constraints, the objective is to formulate a feasible decision rule to maximize total expected return, based on a series of reward functions with capacity decisions x_t and uncertainty $\xi^{[t]}$ as inputs. The objective for the capacity expansion model with flexibility decision rules is captured by (3.5) *s.t.* (3.2)-(3.4) and (3.6)-(3.8).

$$\begin{aligned} Max\ ENPV = \frac{1}{S_s}\sum_{s=1}^{S}(-\mathcal{C}^0(\theta_s^0) + \sum_{t=1}^{T}\left(\frac{1}{1+\lambda}\right)^t [\mathcal{R}_s^t\left(\mathcal{F}_s^t\left(\xi_s^{[t]}, \theta_s^{t-1}\right), \xi_s^t\right) \\ - \mathcal{C}_s^t\left(\mathcal{F}_s^t\left(\xi_s^{[t]}, \theta_s^{t-1}\right), \xi_s^t\right) - \mathcal{H}_s^t(\mathcal{F}_s^t\left(\xi_s^{[t]}, \theta_s^{t-1}\right), \theta_s^{t-1})]) \end{aligned} \tag{3.5}$$

$$\theta_s^t \in \Theta^t \tag{3.6}$$

$$\mathcal{F}_s^t\left(\xi_s^{[t]}, \theta_s^{t-1}\right) = \theta_s^t \tag{3.7}$$

$$\mathcal{H}_s^t(\mathcal{F}_s^t\left(\xi_s^{[t]}, \theta_s^{t-1}\right), \theta_s^{t-1}) = \begin{cases} K(\theta_s^t - \theta_s^{t-1})^\alpha & if \quad \theta_s^t > \theta_s^{t-1} \\ 0 & otherwise \end{cases} \tag{3.8}$$

Note that the reward and cost functions presented in (3.5) differ from the ones (3.1) as they are influenced by the decision rule policy $\mathcal{F}$. The capacity decision made in each period t for scenario s and component i (i.e., ${\theta^t}_s$) can be found from (3.7). The form of $\mathcal{F}$ varies for different problems, and a vector of parameters $\emptyset$ is used to characterize it, with the decision rule represented here as $\mathcal{F}_\emptyset$. Finally, (3.8) gives the general expression for the expansion cost function $\mathcal{H}$, determining the expense associated with each capacity expansion decision, based on the problem at hand. Letting $R(\mathcal{F}_\emptyset\ (\xi), \xi)$ denote the total accumulated reward based on decision rule function and uncertainty realizations, the problem of choosing an optimal decision rule vector is defined by (3.9) below. In practice, $\mathcal{F}_\emptyset$ is restricted to be a subset of X in order to tractably obtain a closed form solution[10].

$$\max_{\mathcal{F}_\emptyset}\left\{\mathbb{E}[\ R(\mathcal{F}_\emptyset\ (\xi), \xi)] = \frac{1}{S_s}\sum_{s=1}^{S}\sum_{t=0}^{T} r_t\left(\mathcal{F}_\emptyset\ \ ({\xi_{[t]}}^s), {\xi_t}^s\right)\right\} \tag{3.9}$$

3.4.3 Capacity Expansion with Deep Reinforcement Learning

A DRL based approach is used to evaluate and optimize an implicit form of *decision rules* to act as a triggering signal for making a particular flexibility decision during system operation, as captured by a policy π within a DNN. The major difference compared to alternative methods discussed in 2.1 or the formulation presented in 3.4.2 above is that π can be characterized by a much larger set of parameters, and thus actions need not be restricted to a static pre-determined form. The primary steps required to formulate an engineering system design and planning optimisation under uncertainty through the DRL Flexibility Analysis framework were described in detail in 3.2.

More generally, it is assumed that the condition of the engineering system for any $t \in T$ can be defined by a state observation, $s^t \in \mathbb{X} \subseteq \mathbb{R}^{n_x}$, where s^t captures the installed capacities for each component i, in addition to other factors. Based on the information given to stakeholders at each evaluation time step, they can sequentially select a system evolution decision, $a^t \in \mathbb{A} \subseteq \mathbb{Z}^{n_a}$, changing the capacity of a component if deemed necessary. Considering the targeted objective, a reward function, R: $\mathbb{X} \times \mathbb{A} \times \mathbb{X} \rightarrow \mathbb{R}$, can be defined which captures the impact of different available actions over time. This is based on a series of periodic revenue $\mathcal{R}^t(s^t,\ \xi^t)$ and cost $\mathcal{C}^t(s^t,\ \xi^t)$ functions, approximated through environment interactions, as well as the *deterministic* expansion cost function $\mathcal{H}^t(a^t, s^t)$ which includes a zero-stage decision on starting configuration.

The engineering system state is then updated as a function of the previous state, action selected, and realisations of stochastic variables. As such, the system is identified to hold the Markov Property, and can be formulated within the MDP framework. The objective, as for the inflexible baseline or flexibility decision rule approach, is generally the maximization of ENPV through the development of a system evolution policy, $\pi: \mathbb{X} \rightarrow \mathbb{A}$. Note that a^t is then obtained from π by using s^t as an input, as shown in Figure 2.3 and Figure 3.5. The generic model for DRL based engineering system flexible capacity planning is defined according to (3.10), and can be implemented on the same scenarios S_s as in (3.1) or (3.5) above. Furthermore, as $\pi(\emptyset,\cdot)$ is approximately captured by a NN with parameters $\emptyset \in \mathbb{R}^{n_\emptyset}$, the capacity expansion problem simplifies to obtaining optimal network parameters $\emptyset^*$ [134].

$$\mathcal{P}(\pi) := \begin{cases} \max_{\pi} \mathbb{E}_{\pi}[NPV] \\ s.t. \\ NPV_s = \sum_{t=0}^{T} \left(\frac{1}{1+\lambda}\right)^t {R^t}_s \\ {R^t}_s = \mathcal{R}^t(s^t, \xi_s^t) - \mathcal{C}^t(s^t, \xi_s^t) - \mathcal{H}^t(a^t, s^t) \\ s^{t+1} = f(s^t, a^t, \xi_s^t) \\ a^t = \pi(s^t) \\ \forall\, t \in T\,, s \in S \end{cases} \tag{3.10}$$

3.4.4 Uncertainty Simulation

The development of stochastic models to account for uncertain variables is a very broad field, particularly when applied to engineering systems with a high perceived risk level. Several potential options exist such as Bayesian theory, Dempster-Shafer, jump diffusion models, mean reversion, scenario planning, and more advanced statistical approaches such as ANN based forecasting[6,210,211]. Determining the appropriate approach for each variable, however, is an extremely complex topic onto its own, and a whole thesis could be written just on optimizing the trade-offs among alternative stochastic modelling methods. In order to focus on the objectives of this thesis, therefore, Geometric Brownian Motion (GBM) models are developed to estimate stochastic evolutions of most case study variables simulated, unless stated otherwise. This form of stochastic model is chosen as it can represent, for any variable x, both overall long-term trends (if any), and the potential for tail event or low probability scenarios. This is deemed highly compatible with most of the uncertainties considered in this work, as found in previous studies[62,77,212,213]. It should be noted, however, that if available data were to suggest otherwise, alternative, and more complex stochastic model formulations should be explored outside this proof-of-concept work- as discussed in 3.1.3. Mathematically, this GBM involves drift μ_x and volatility σ_x variables, respectively. dW_t models the random Wiener process, or stochastic deviations from the mean as sampled from a normal distribution, and dt the time resolution implemented. These values, typically based on historical data, allow one to fine-tune the stochasticity

of the model, while not limiting the potential total evolution scenarios [214-216]. They can then be used to simulate uncertainty for any variable X as shown by (3.11)-(3.12) below[134].

$$dX^t = \mu_x X^t\, dt + \sigma_x X^t dW_t \tag{3.11}$$

$$X^t = X^0 \exp(\mu_x - {\sigma_x}^2/_2)t + \sigma_x W_t \tag{3.12}$$

3.4.5 The Value of Flexibility

The total value of flexibility ${VoF^z}_{Total}$ for any engineering system considered, on any metric z, is captured by (3.13). In essence, it comes from the combination of tactical system planning flexibility ${VoF^z}_{Strategic}$, and the added value enabled by operational flexibility possibilities ${VoF^z}_{Operational}$. As shown in (3.14), it is thus estimated as the difference between the expected value for the best flexible solution ${EV_{Flexible}}^z$ and the best performing inflexible benchmark ${EV_{BestBaseline}}^z$. For $z = ENPV$, as formulated for all methods in 3.4, ${VoF^{ENPV}}_{Total}$ represents the upper bound that should be paid to embed said flexibility in an engineering system design. This helps inform better decision-making since the proposed approach helps improve assessments on the value of flexibility, which can then be used in design analysis to determine whether the cost of embedding it are worth it.

$${VoF^z}_{Total} = {VoF^z}_{Strategic} + {VoF^z}_{Operational} \tag{3.13}$$

$${VoF^z}_{Total} = {EV_{Flexible}}^z - {EV_{BestBaseline}}^z \tag{3.14}$$

$${VoF^z}_{Operational} = EV^z_{Flex\ Operation} - EV^z_{Inflex\ Operation} \tag{3.15}$$

$$\forall\, z \in Z \tag{3.16}$$

Furthermore, (3.15) is used to decompose the operational flexibility $VoPP^z$ possibilities from the DRL based decision-making flexibility proposed in this work. The value of operational flexibility is thus approximated from the difference in performance for baseline designs $EV^z_{Inflex\ Operation}$ as compared to ones integrating flexible operation $EV^z_{Flex\ Operation}$. While it is recognized the values for these flexibility sources may not be purely additive, they are considered separately in this thesis to determine whether or not the infrastructure investment required to enable the operational flexibility developments discussed in Chapter 6 are worthwhile in the targeted context. Decomposing relative performance improvements by source may also be beneficial for various policy-making objectives. It gives relevant decision makers important quantitative insights across different metrics Z. These can then be used to help develop market support instruments or incentive programmes, with the aim of maximizing added value from flexibility implemented in different manners. The specifics on how these metrics are calculated for each case study are further detailed in their respective sections.

Chapter 4

4 Methodological Validation: Proof of Concept Infrastructure System

This chapter is focused on one essential question for validating the novel proposed approach:

- RQ3: Is it possible for DRL to outperform existing methods in Flexibility in Design on a well-known simple proof of concept case study?

This question is tackled using the simple infrastructure system first introduced as a motivating example in 3.1.2. This is a foundational problem in ROA and Flexibility in Design which has been benchmarked using all commonly implemented approaches, representing an important starting point for the validation of the proposed DRL based method. It should be noted that this was the first case study assessed as part of this thesis, and as such the methods and analysis presented are simpler than in the following chapters. Nonetheless, the results suggest significant potential added value through the proposed approach, motivating further validation and development on more complex case studies. The full open source code to reproduce all results presented in this chapter can be accessed here[217].

4.1 Case Study Overview

In order to test the potential of DRL to produce value enhancing DRs in systems engineering design, the approach is first tested on a simple infrastructure system. The complete strategic design problem is taken from the development of a multi-story parking facility in the UK, where the project managers and system designers regarded the growth in customers/ parking demand for the adjacent mall to be highly uncertain as it had not yet been built. As a result, the objective was to evaluate whether a more flexible design which could allow capacity increases over time could increase expected value for the project compared to a stochastically optimal fixed design. The garage presents a capacity of 200 cars per level, a maximum allowable height of 9 floors, and the possibility to yearly decide whether to expand capacity by 1, 2,or 3 floors – as shown in Figure 4.1 (i) [184]. This problem was chosen as it is an often-used pedagogical case to teach the ideas of flexibility in design due to its relative simplicity with few numbers of design variables involved. It presents tensions between maximizing economies of scale (EoS), an approach often used in system design to deal with uncertainty which can result in excessive early capacity commitment, and smaller modular deployments which may lead to higher total investment cost, albeit providing valuable adaptability, as shown in Figure 4.1 (ii). In short, the graphic captures how a rigid design exploiting EoS does not necessarily perform better than a flexible modular design with no EoS when considering uncertainty in key variables.

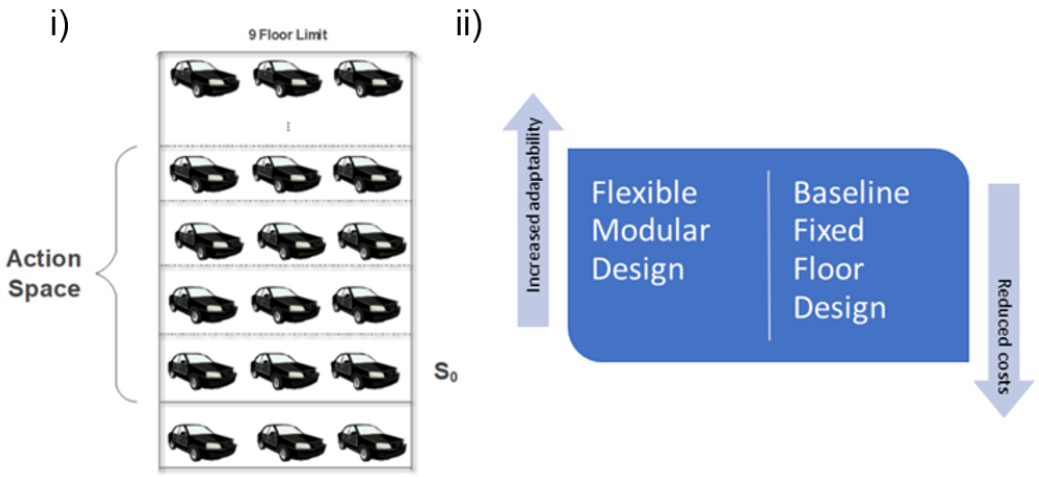


Figure 4.1: i) Example multi-story garage; ii) Overview of tensions involved in design process.

4.1.1 Mathematical Model and Assumptions

A discounted cash-flow (DCF) model is built to represent the relevant variables in this infrastructure design problem. The objective is to maximize the Expected Net Present Value (ENPV) of the project over a 20-year operational life T for all scenarios $s \in S$, as shown in (4.1), where λ represents the 12% discount rate used. The general formulation for the infrastructure system model is defined by (4.2)-(4.6). The net cash flow per period NCF_t is given by (4.2) as the difference between annual revenues R_t and costs C_t. The expression for R_t shows that the maximum revenue in each period t is capped by the installed capacity k_t or periodic demand D_t. An annual price p of USD 10,000 is charged per space used. Total yearly costs C_t are instead found from the total operational costs $k_t c_r$, land leasing cost c_l, and expansion costs c_{exp}. Land leasing costs are deemed fixed at USD 3.6 Million /year, with an average annual operating cost of USD 2,000 for each space *available* (even if not used). The total installed capacity is given by (4.3) as the sum of initial one and all previous expansion decisions. In the inflexible case, $f_t = 0 \; \forall \; t$ since no expansion decisions are possible, while in the flexible case $f_t \neq 0$ to allow expansion as demand increases over time. The initial investment C_0 is determined from (4.4) as the sum of construction cc_0 and land leasing costs. Initial construction costs cc_0 are calculated as $cc_0 = n_0 f_0 c_0$ for the first two levels built, where f_0 represents the starting number of floors, n_0 the 200-car capacity per floor, and c_0 the cost of pre-cast construction at USD 16000/space. Past the first two floors, costs are assumed to grow at a rate of 10% per level. To capture technical limitations of the project, (4.5) is a constraint on the allowable capacity for the garage, bounded by a maximum number of 9 floors f_{max}. Initial deterministic projections for demand over time D_t are modelled using (4.6). Here D_f is the final demand in year 20, while α and β are variables computed to capture the exponential demand function predicted at project start, using additional forecasted demand by years 10 and 20. Deterministic assumptions are that the demand at project start D_1 consists of 750

parking spaces, with demand for an extra 750 and 250 spaces by year 10 and 20, respectively. These assumptions result in $\alpha = 1000$ and $\beta = 0.15$ in the deterministic demand function from (4.6).

$$ENPV = \frac{1}{S}\sum_{s=1}^{S}\sum_{t=0}^{T}\left(\frac{1}{1+\lambda}\right)^{t} NCF_t \tag{4.1}$$

$$NCF_t = \min(D_t, k_t)\, p \; - \; k_t c_r - \; c_l - c_{exp} = \; R_t - \; C_t \tag{4.2}$$

$$k_t = f_0 n_0 + n_0 \sum_{t=1}^{T} f_t \tag{4.3}$$

$$C_0 = cc_0 + c_l \tag{4.4}$$

$$k_t < n_0 f_{max}\,, t > 0 \tag{4.5}$$

$$D_t = D_f \alpha e^{-\beta t} \tag{4.6}$$

In practice, however, the evolution of demand is highly uncertain due to the potential development of a nearby mall. Consequently, the stochastic demand model given by (4.7) is formulated as GBM around an S-trend curve, in line with previous works on this case study, a slight modification to the general formulation from (3.11)-(3.12). Here, D_t is still calculated according to (4.6), allowing out of sample testing as detailed in 4.4. Additional demand for years 10 and 20 is instead sampled from a uniform distribution, with realized parking demand ± 50% from the original projections. Stochastic inter annual demand growth g_t is found from (4.8), where g_p is the projected deterministic growth rate. The yearly increments for simulation dt and the respective estimated annual volatility σ of 15% can then be used to generate a potentially infinite number of *i.i.d.* scenarios of demand evolution.

$${D_{t+1}}^{S} = g_t(1 + \; D_t) \tag{4.7}$$

$$g_t = g_p D_t + \; \sigma dW_t \sqrt{dt} \tag{4.8}$$

The introduction of demand profile uncertainty is captured by Figure 4.2 below, highlighting the potentially large variability in market operating conditions encountered by the engineering system. Low, medium, and high growth stochastic scenarios are included to give some intuition on the need for adaptability in this context. As can be seen, while the deterministic forecast assumed constant and predictable growth, the reality changes significantly when introducing deep uncertainty. System designs produced only using deterministic data, therefore, are likely to perform poorly in a lot of scenarios -driven by either enhanced downside exposure or reduced ability to capture upside profits. Similarly, stochastically optimal inflexible baselines, while performing better under uncertainty than deterministically optimal ones, provide limited adaptability to different scenario realizations.

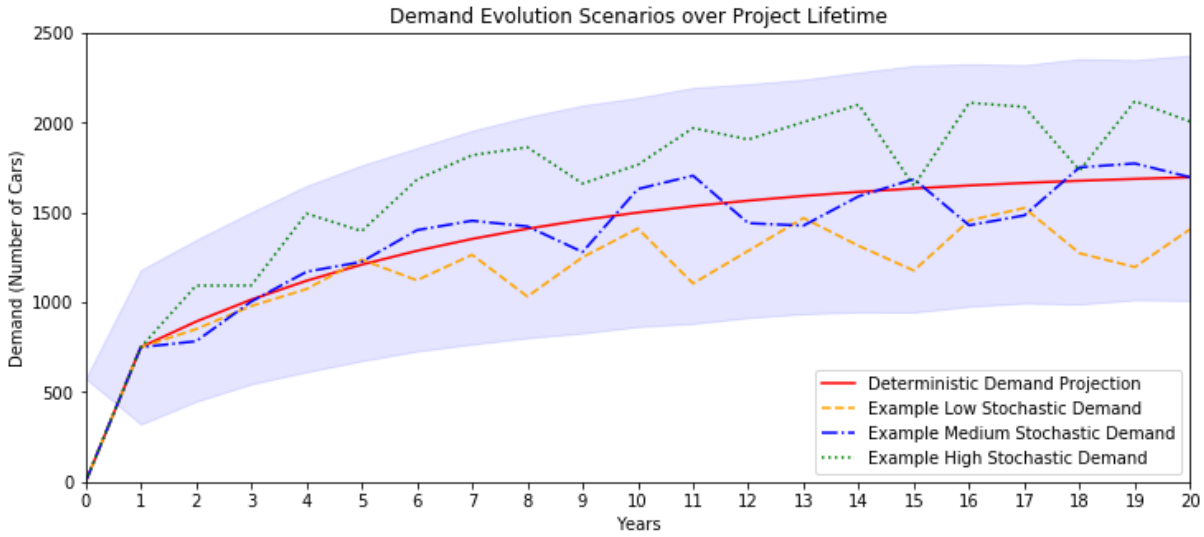


Figure 4.2: Comparison of deterministic and stochastic simulated demand evolutions. The shaded blue area represents the 95% confidence interval for the stochastic demand over 2000 simulated scenarios.

Please refer to 11.1 for the complete list of assumptions and parameters used in the model detailed in this section.

4.2 Benchmarking

This section presents the most extensive benchmarking of DRL agent performance in this thesis as it was deemed a core issue to methodological validation. In light of the above, the DRL based design solution is benchmarked in terms of expected economic performance under uncertainty against:

A. Stochastically optimal inflexible baseline.
B. Flexibility Decision Rules.
C. Dynamic Programming Sample Average Approximation (DP SAA).

Design alternative A is included to evaluate trade-offs among economies of scale and the inclusion of flexibility in the system. Design alternative B yields interesting insights on a simpler and more intuitive method to integrate adaptability, however, can sometimes yield slightly sub-optimal solutions due to the pre-restricted form they must take, which are not applicable across all scenarios. Finally, design alternative C represents a state-of-the-art method for improving expected performance in decision-making under uncertainty, although criticized as less interpretable than the decision rule approach. Together, they represent the primary alternatives available for engineering systems design and operation and can confidently validate the comparative performance of the proposed methodology on the foundational case study for Flexibility in Design. Unless stated otherwise, all solutions are obtained based on the same sample of 2000 *i.i.d.* stochastic scenarios for improved comparability.

4.2.1 Inflexible Baseline

This section introduces the generic formulation of the rigid capacity planning model under uncertainty to determine the stochastically optimal number of fixed floors to build. The optimisation problem, not considering expansion decisions, simplifies to (4.9). The reward function per period $r_t{}^s(f_0, \xi_t{}^s)$ is based on the uncertainty realizations in scenario s and is equivalent to the NCF for year t defined in (4.2). No expansion costs are considered in this design.

$$\max_{f_0} \frac{1}{S}\sum_{s=1}^{S}\sum_{t=0}^{T} r_t{}^s\,(f_0, \xi_t{}^s) = \max_{f_0} \frac{1}{S}\sum_{s=1}^{S}\sum_{t=0}^{T} \left(\frac{1}{1+\lambda}\right)^t NCF_t \tag{4.9}$$

s.t. (4.2)-(4.5)

This problem is fairly simple to optimize due to the presence of only one design variable f_0 with a small feasible design space involved, even if presenting a highly non-linear and disconnected objective function. The Differential Evolution (DE) method from[218] is applied to the optimization, quickly converging to the optimal solution of building **6 floors** at project start, in line with previous results[184]. Past this point, it is unlikely to recover investments made in infrastructure, while building less floors leads to excess demand not taken advantage of in many scenarios. The expected performance of this design under uncertainty is presented in more detail in 4.4.

4.2.2 Flexibility Decision Rules

The potential to embed flexibility within this engineering system is first evaluated using the Flexibility DRs approach, optimized via the objective presented in (3.9). In this proof-of-concept study, a design vector is initially defined capturing the most impactful decision-making possibilities by the agent. This is then consulted at every decision stage to ultimately determine the capacity expansion schedule. The possible values and decision rules analyzed as part of the design vector are summarized in Table 2, providing intuitive triggering signals for decision-makers during operation.

Table 2: Decision rules and subset of values considered in the design vector for the multi-story parking garage case study.

Decision Rule Variable	**Explanation**	**Levels Evaluated**
$\boldsymbol{a_{1-4}}$	Allow expansion in years 1-4	Yes-No (1-0)
$\boldsymbol{a_{9-12}}$	Allow expansion in years 9-12	Yes-No (1-0)
$\boldsymbol{a_{17-20}}$	Allow expansion in years 17-20	Yes-No (1-0)
$\boldsymbol{d_r}$	Previous years with excess demand before expansion	2-4

f_t	Floors expanded by on expansion years	1-3
f_0	Starting number of floors	1-9

The decision rules considered in this section are classified as conditional-go, meaning they are formulated via a set of IF-THEN-ELSE statements based on the above variables. Their logic is graphically captured in Figure 4.3. It shows how historical data on system configuration (starting from f_0) and uncertain variable realization is first used by **Procedure 1** to determine whether there has been unmet demand in the past d_r years. **Procedure 2** then checks whether the binary decision variable constraints allow expansion for that specific year (i.e. a_{1-4}, a_{9-12} , and a_{17-20} are equal to 1 for t is equal to 1-4, 9-12 and 17-20 respectively). Finally, given both Procedure 1 and Procedure 2 have been met, the expansion magnitude (i.e. by f_t levels) is executed in **Procedure 3**.

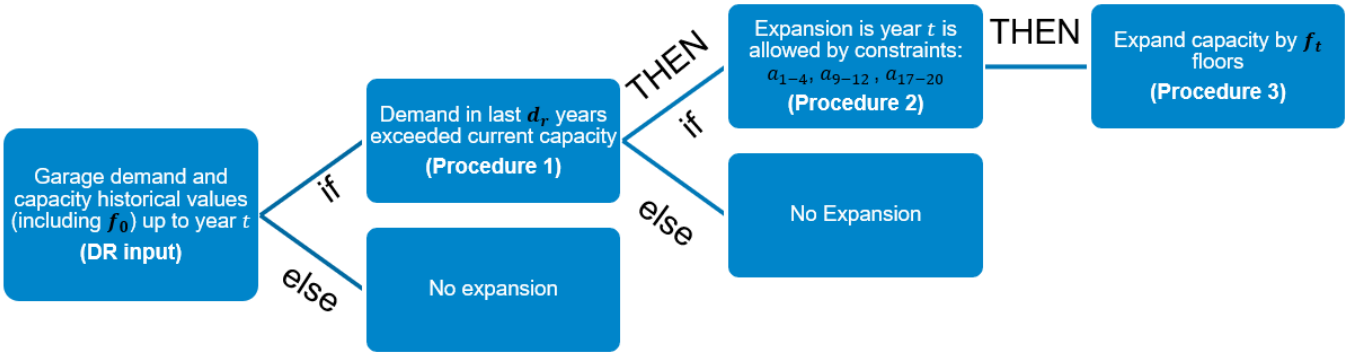


Figure 4.3: Visualization of flexibility decision rules procedures for multi-story garage case study. Decision rule variables optimised are highlighted in black for improved graphical clarity.

These decision rules are optimised using the formulated design vector and the DE algorithm for a more comparative evaluation. They also include the only design variable considered in 4.2.1, f_0, which is the starting number of floors built at year $t = 0$. Several runs suggest solutions convergence to the likely global optimum fairly quickly, although hyperparameter tuning remains important. The final optimal design vector solution found is in (4.10), suggesting a more dynamic solution to the baseline as it is able to consider both EoS and modular deployment by enhancing system flexibility.

$$[\, a_{1-4}, a_{9-12}\,, a_{17-20}, d_r, f_t, f_0\,]^* = [\,1, 1, 0, 2, 2, 4] \tag{4.10}$$

4.2.3 Dynamic Programming Sample Average Approximation

The DP SAA approach is implemented as presented in [219] including sampling strategy, as it is well suited to solve Monte-Carlo simulation based discrete decision analysis problems such as the one evaluated here. The SAA problem is generally defined by (4.11) below as:

$$z_n{}^* = \max_{x \in X} \frac{1}{S} \sum_{s=1}^{S} R(\mathrm{x}, \xi^s) \tag{4.11}$$

where $R(\mathrm{x}, \xi)$ denotes the accumulated reward in one scenario s, or the NPV. In this method, samples are repeatedly drawn from the distribution of uncertainty and the resultant average sample deterministic optimization problem is solved with backward induction. Each sample may lead to a different solution, however if the sample size is large enough, the sample objective function value will approximate the true one[53]. The true problem and expected reward are thus approximated via the sample average function from (4.11), yielding $z_n{}^*$ as an estimator of the true optimal value and $x_n{}^*$ as an estimator of the true optimal policy. An optimality gap of 2% is used as a stopping criterion. The primary difference between this and flexibility DR approach is that in the SAA approach, the optimal policy or set of actions does not have to stay constant across all time periods and scenarios. This allows more valuable decision-making to be considered by the system, but at the cost of a less usable policy and a greater computational expense. Contrary to 4.2.1 and 4.2.2, in fact, no intuitive explanation of the mathematical solution form can be presented, as has been previously criticized.

4.3 DRL Implementation

4.3.1 DRL Flexibility Analysis Formulation

The DRL formulation of the capacity planning problem detailed in 3.4.3 and applied to the infrastructure case study from this section is developed as an Open AI Gym environment. The Open AI gym format is followed as it is by far the most prevalent in the RL community, providing a standardized and reproducible way to assess environment feasibility and agent performance[161]. Furthermore, all major DRL packages and libraries mentioned in 2.3.3 have been designed to be compatible with Open AI gym environments, allowing for simple and fast experimentation among different algorithms. The different steps for DRL Flexibility analysis are further detailed below.

1. **Environment Formulation:**
    i. State Space: In this proof-of-concept case study, developing a representative DRL environment is fairly straightforward due to the limited number of variables involved. The state space s^t consists of a *Box* observation set capturing system capacity θ^t and uncertain variable realization ξ^t (in this case only demand D^t) for year t, normalised by their lower and upper boundaries, respectively. Note the difference compared to the alternative methods presented in 4.2.2 and 4.2.3, where the information up to and including the time of decision is needed to determine the exercise of a flexibility strategy. In order to address this, time awareness is also included in the agent observations based on previous works suggesting significant training performance improvements in finite-horizon problems[220]. This approach does not violate nonanticipativity constraints as remaining project expected duration $T - t$ should be

part of the description of the environment to maintain the Markov Property. The complete observation set is thus defined according to (4.12).

$$s^t = [\theta^t, D^t, T - t] \tag{4.12}$$

ii. Reward Function: Similarly, the form of the reward signals is determined by the nature of the problem, as it considers investment purely from a financial point of view. As such, the reward for each timestep t is defined as the NCF for that particular period computed according to (4.2), ultimately leading to the generic ENPV maximization problem shown in (3.10). Furthermore, an imaginary reward penalty of -USD 1 is included in time steps where the DRL agent violates problem constraints, namely the maximum allowable capacity in this case given by (4.5). This is a simple albeit commonly used approach to let the DRL agent learn the forms of the constraints implicitly through interactions, while not increasing the solution procedure complexity of the policy optimisation process. It is an arbitrary value whose effectiveness for different goals is typically determined through a mix of intuition and a trial-and-error learning process. The reward signal r^t recorded in each year t is thus computed as shown by (4.13) below.

$$r^t = \begin{cases} NCF_t & if\ \theta^t \leq \theta^{max} \\ -1 & otherwise \end{cases} \tag{4.13}$$

2. **Uncertainty Quantification:** The characterization and quantification of uncertainty for this problem was already detailed in 4.1.1. It is not modified for the DRL implementation to allow a more comparative evaluation to the alternative solution towards the validation of the proposed methodology. One important difference, however, is that the DRL agent can tractably handle a much larger number of scenarios than alternative methods, helping to enhance generalization of designs and strategies under uncertainty. These scenarios are generated according to (4.7) and (4.8), using the assumptions listed in 4.1.1. The realization of uncertainty over time and across scenarios is captured by variable D^t in the state space.
3. **Action Space Creation:** Given the methodological validation objectives driving the analysis of this case study, only the same level of flexibility enablers presented in 4.2.2 and 4.2.3 are considered in the development of the DRL agent action selection possibilities. A *Discrete* action space is thus formulated to account for the different decision options over time enabled by the system. For any year t, the agent may thus choose to expand by 0,1,2, or 3 floors. Note for the $t = 0$ decision on starting capacity f_0, the action space is expanded to include 1-9

levels as described in Table 2, allowing a more comparative evaluation. Past the zero-stage investment decision, the discrete action space is captured by (4.14) below.

$$a^t = \begin{cases} 0 \rightarrow & No\ Expansion \\ 1 \rightarrow & Expand\ by\ 1\ floor \\ 2 \rightarrow & Expand\ by\ 2\ floors \\ 3 \rightarrow & Expand\ by\ 3\ floors \end{cases} \quad \text{for t} > 0 \tag{4.14}$$

4.3.2 Algorithmic Approach

Given this chapter is focused on validating the proposed methodology in a case study with relatively low dimensionality and a discrete action space, the capacity expansion problem under uncertainty is tackled using simpler value-based methods. More specifically, the Double Deep Q-Network (DDQN) agent with prioritized experience replay (PER) was implemented[166]. This algorithm leverages the approximation power of neural networks to iteratively improve a complex representation of the value function through sequential agent-environment interactions. DDQN works by simultaneously updating network parameters based on reward signals received, with the objective of minimizing error between the true and approximate value of state-action pairs. The use of separate networks for value estimation and action selection allows for improvements in performance stability and helps address optimistic Q-value bias[139]. Furthermore, the original DDQN is augmented by focusing on transitions or episodes which yield the largest changes in agent performance, rather than uniformly sampling from past experiences[141]. The DDQN agent is thus exposed to replays of important transitions more often, helping to increase learning efficiency and resilience of developed policy.

Action selection in such a complex stochastic environment, however, introduces the important notion of *exploration vs exploitation* for the learning agent. This is equivalent to the compromise between continuing to try new actions which could result in increased return or sticking with current actions which are known to have a high value. The common ε-greedy behavior policy shown in (4.15) is chosen to address this issue. It is equivalent to taking the highest valued action with a probability of 1-ε, while uniformly random sampling from the action space with a probability of ε to enhance exploration. To balance this, ε is initially set to 0.5 here, with a linear decay schedule. The probability $P(i)$ of each memory being sampled is proportional to its training-loss adjusted TD error, as shown in (4.16). The exponent α determines the extent to which prioritization is used, where $\alpha = 0$ represents the uniform sampling case, and $p_i > 0$ is the priority assigned to transition i .β quantifies the extent to which we want to compensate for the bias error (0-1) and k is the minibatch size. β is normally annealed to 1 during the training process as unbiased updates become more important.

$$\pi\ (a_t|s_t) = \begin{cases} \arg max_{a_t} Q(s_t, a_t)\ , & w.p.\ 1-\varepsilon \\ random\ action\ , & w.p.\ \ \varepsilon \end{cases} \tag{4.15}$$

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \tag{4.16}$$

The full pseudocode for DDQN with PER applied to this case study is given by Algorithm 1 below.

Algorithm 1: Double DQN with prioritized experience replay applied to garage case study, adapted from [166].

Initialize Q network parameters θ, minibatch k, step-size η, replay period K and size N, exponents α and β, budget T, initial exploration epsilon ε, M training episodes

Initialize replay memory $H = \emptyset, \Delta = 0, p_1 = 1$

For M episodes do:

  Select starting number of floors (f_0) based on initial demand (S_0) and current policy $A_0 \sim \pi_\theta(S_0)$

  for $t = 1$ to T do:

    Take random or highest valued capacity expansion decision (A_t) according to (4.15)

    Observe transition to next state (S_t) and resulting yearly NCF in *t* (R_t)

    Store transitions ($S_{t-1}, A_{t-1}, R_t, S_t$) in H with initial maximal priority

    if replay period K reached **then**

      for $j = 1$ to k **do**

        Sample transition $j \sim P(j) = p_j^\alpha \,/\, \sum_i p_i^\alpha$

        Compute importance sampling weight $w_j = (N \cdot P(j))^{-\beta} / max_i w_i$

        Compute TD error $\delta_j = R_j + \gamma_j Q_{target}(S_j, \arg max_{a_j} Q(S_j, a_j)) - Q(S_{j-1}, A_{j-1})$

        Update transition priority $p_j \leftarrow |\delta_j|$

        Accumulate weight change $\Delta \leftarrow \Delta + w_j \cdot \delta_j \cdot \nabla_\theta Q(S_{j-1}, A_{j-1})$

      end for

      Update weights $\theta \leftarrow \theta + \eta \cdot \Delta$, reset $\Delta = 0$

      Optimize Q-network parameters and copy onto target network periodically $\theta_{target} \leftarrow \theta$

    end if

    Choose action $A_t \sim \pi_\theta(S_t)$ or expansion level decision based on current parking demand

  end for

The logical progression of the proposed approach based on DRL Flexibility Analysis and Algorithm 1 is graphically summarized by Figure 4.4 to provide some intuition on both the learning and decision-making process. It presents several important differences to the logic for the FDR approach depicted in Figure 4.3. Generally, the state observation (built capacity and parking demand here) in any given year is used as an input into the NN parametrized value function Q_θ, not requiring a history of previous observations. In fact, Q_θ estimates the expected return for each action available from any

input state (for example limited by constraints on maximum capacity deployed). The DDQN agent then selects an action based on (4.15) and executes it in the environment model, leading to an updated system state and NCF (or reward) in the following year. At first, the agent's actions tend to be highly random and exploratory, followed by convergence to highest valued ones over the learning process. By efficiently evaluating different regions of the decision-making space, the agent generates data to provide better guidance on approximate value function updates, and in turn action selection.

The final optimal decision-making policy $\pi^*(a_t|s_t)$ can then be implicitly derived from the final parametrized state action value function, providing a deterministic mapping of actions from state observations. It can be directly used by system operators to determine capacity expansion decisions under uncertainty more dynamically than with alternative methods, by taking the maximum over all actions available from a system configuration. As shown in Figure 4.3, a simple fully connected NN architecture (or MLP) is implemented for this proof-of-concept case study, with input and output layers matching observation and action space dimensionality, respectively. Additionally, two hidden layers with 16 neurons each are used for mapping input features to action selection, balancing expressibility of the estimated value function with speed, expense, and difficulty of training. Note that as discussed above, the target network used for policy updates presents the same architecture as the value network (albeit using the older set of weights θ^-).

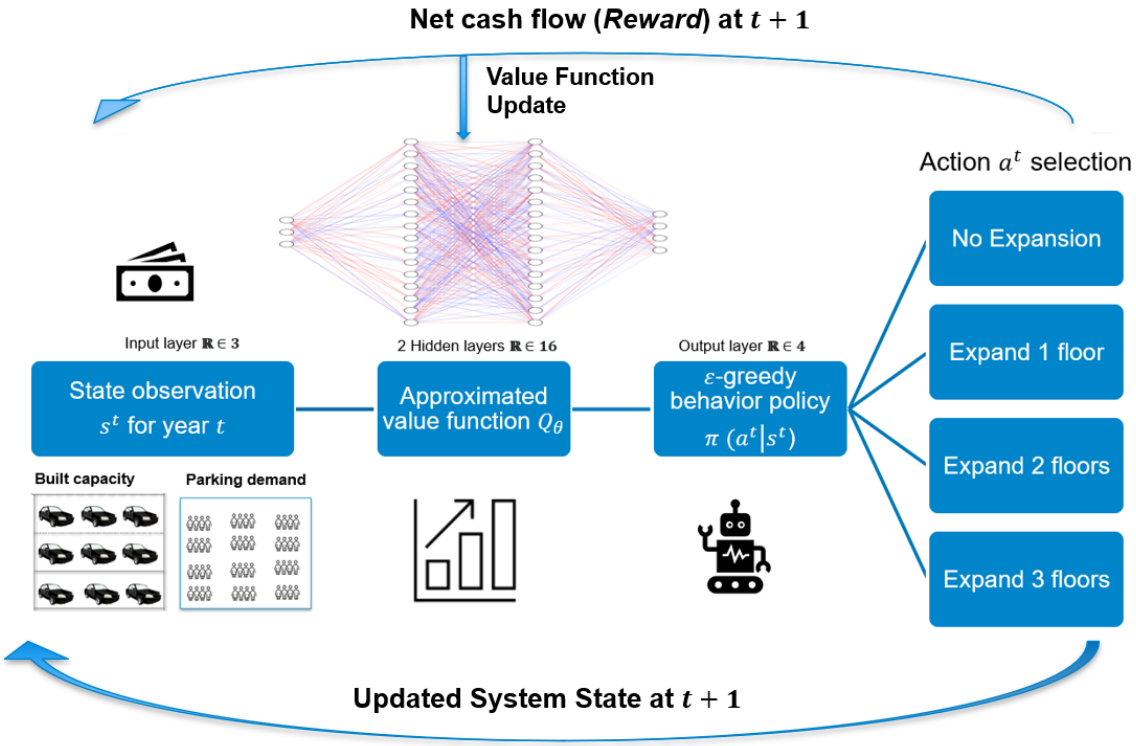


Figure 4.4: Graphical overview of DRL agent decision-making process on garage case study using Algorithm 1.

4.3.3 Training and Testing

The DDQN agent described in the former section is trained using a budget of 500,000 yearly decision timesteps, equivalent to 25,000 simulated full 20-year episodes. As performed on a standard laptop Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz, the complete training process for a single agent takes approximately 1-2 hours. All training scenarios are stochastically generated to be *i.i.d* from each other, thus leading to an *infinite* amount of potential scenario evolutions and state-action pairs. This was motivated by the need to develop a more generalizable adaptability strategy under different realizations of uncertainty. No exactly equal trajectories are thus possible inside the environment, and training process is fully non-anticipative, with all operational decisions based purely on recorded observations at a particular timestep[134]. Hyperparameters in this initial implementation are found through a combination of grid search and intuition on the formulation of the environment (please refer to 11.2 for a complete list of hyperparameters used). Changes to them from default ones are observed to substantially impact both training efficiency and final agent performance, motivating more advanced approaches for hyperparameter tuning following the methodological validation stage. All NN parameters θ are updated using the Adam optimizer here[124].

The complete learning curve for the DRL agent on this case study is shown in Figure 4.5, averaging results from 5 separate training runs using different random initialization seeds. Note that only the best training run is used for relative performance analysis in 4.4. Evaluations of sum of rewards were conducted every 100 episodes, using a greedy policy (i.e. $\varepsilon = 0$) for 5 episodes. The results suggest that the proposed DRL approach is able to learn effective policies in all runs, albeit experiencing significant variance during the training process. This can ultimately lead to substantial differences in final trained agent policies, with only the best performing run compared to the benchmarks in 4.4. In the initial training episodes (i.e. 0 to 150,000 time steps), the DDQN agent focuses on exploration of the decision-making space and often is penalized for constraint violation – resulting in overall very negative ENPV values. After this, however, the DRL agent seems to gradually exploit data from environment interactions to improve the value function approximation, leading to more dynamic and reward maximizing action selection. According to Figure 4.5, the policy generally converges past around 400,000 time steps, aligned with more limited exploration of the environment in the final stages of training.

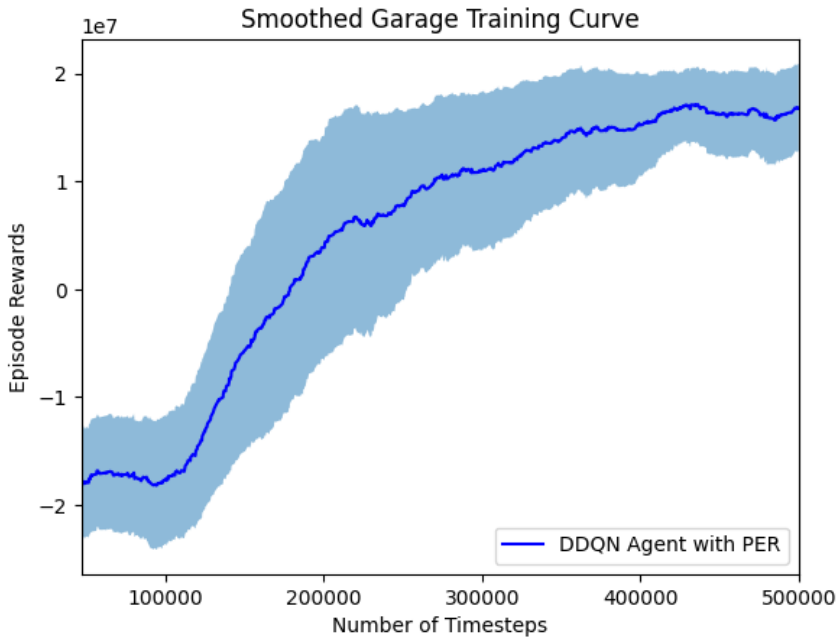


Figure 4.5: Learning curve for DDQN agent on garage capacity expansion case study, where the y-axis represents the ENPV (USD) obtained in one episode (20 year project life) and the x-axis the number of training decision time-steps. The solid lines show the smoothed averages using a window size of 50 episodes. The shaded area shows the standard deviations among different training runs.

The arbitrary reward penalty in (3.14) for deploying capacity past its maximum level is then removed during the testing/benchmarking phase, as it is not representative of real-world operation. Consequently, the true ENPV obtained by the DRL based flexible capacity expansion strategy is evaluated, as opposed to the return computed via the modified reward function during training. An important distinction is noted here in the optimization of system designs and policies using alternative methods, which only used the same 2,000 fixed *i.i.d.* scenarios to maintain computational tractability. The DRL approach, on the other hand, trains over 25,000 *i.i.d.* scenarios to enhance generalizability. Iteratively training using only the same 2,000 fixed scenarios would have likely yielded much faster convergence but sacrificed generalizability under uncertainty. Nonetheless, for the benchmarking presented in the following section, DRL agent performance is compared using the same set of stochastically generated demand scenarios as implemented in the alternative methods. This means that while trained on the same stochastic model *parameters*, no complete scenario used in testing would have been encountered during training (i.e. *out of sample* testing). The different time-step resolutions implemented for the DRL agent in this case study are summarized by Figure 4.6.



Figure 4.6: Summary of timestep resolutions used in different stages for the proposed DRL approach on the multi-story garage case study.

4.4 Relative Performance Evaluation

4.4.1 Numerical Results

4.4.2 Motivations for Flexibility Based System Design

Ultimately, the methodology proposed in this thesis is strongly based on recognizing flexibility as a value-enhancing paradigm under uncertainty, estimated via (3.14). In the methodological validation stage, however, we performed a preliminary analysis to highlight the motivation for a Flexibility in Design approach in capacity planning problems, even in a deterministic setting. The deterministic projections are thus used as a reference demand scenario for initial analysis, and the VoF is simply included for comparing the value obtained in the following section. The results summarised in Table 3 show that the flexible design solutions significantly outperform the baseline, with a slight edge for the proposed DRL approach, driven by a more dynamic decision-making policy. On one side, the staged deployment solutions are able to better discount investments into the future. On the other, the demand in the deterministic profile exceeds the 6-floor capacity by the end of year 5 already, which the flexible designs are able to capitalize on by progressively increasing capacity to the appropriate level. Designs focused on flexibility, therefore, also present much lower initial investment, decreasing project risk and increasing attractiveness to investors.

Table 3: Economic performance and VoF from the solutions in a deterministic demand scenario only, with all values shown in USD Million units.

Metric	**Inflexible Baseline**	**Flexibility DRs**	**DRL Flexibility Analysis**
Deterministic NPV	10.56	15.05	15.25
$\boldsymbol{VoF}_{\boldsymbol{Total}}$	0	4.49	4.69

4.4.3 Stochastic Benchmarking

The relative performance of the different design alternatives detailed in 4.2 under uncertainty is graphically captured by Figure 4.7. All flexible design solutions outperform the inflexible baseline based on ENPV, downside exposure and ability to capture upside potential. The cumulative distribution functions (CDF) suggests the proposed DRL based approach can yield a more valuable design solution and flexibility strategy than alternative state of the art methods. Given the same level of technical enablers are embedded in all flexible solutions, this result points to the value of improved decision making under uncertainty, rather than simply unexploited technical flexibility. Designs 2 and 3, in fact, present overall very similar performance in this proof-of-concept study.

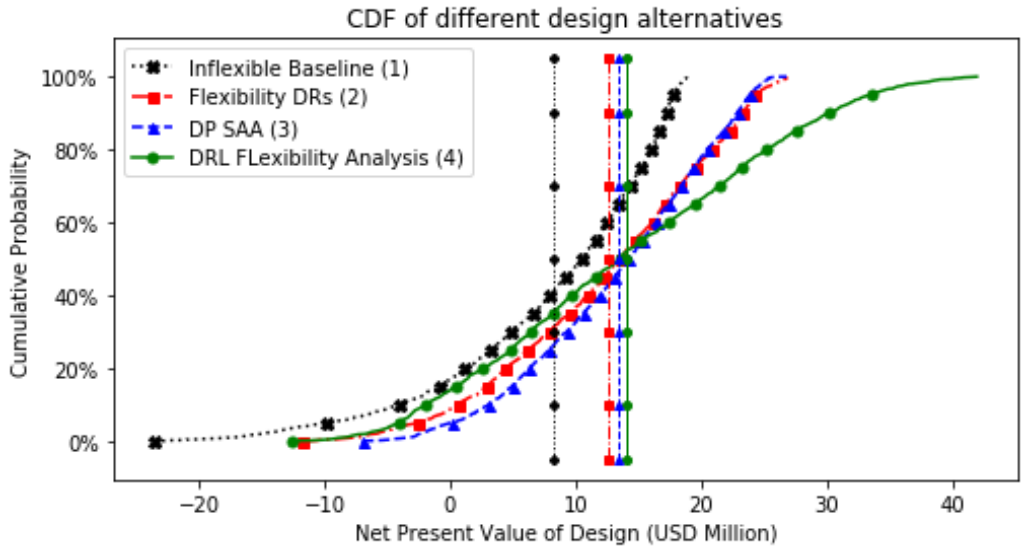


Figure 4.7: Cumulative distribution functions showing lifetime performance of different design alternatives for the garage case study based on 2000 simulated scenarios. Vertical lines represent the ENPV for each of the methods implemented.

Table 4 extends this analysis via a multi-criteria comparative assessment of the various design solutions produced under stochastic demand assumptions. This sort of comparative table is useful as different decision makers may look at different risk factors when evaluating a project, thus helping to give a more holistic overview of system design performance. It highlights the validity and contribution of the proposed DRL approach through this proof-of-concept case study, as it outperforms all benchmark designs on ENPV, VoF, and Value at Gain (VaG). More specifically, ENPV is increased by **18%** and **11%** relative to the Flexibility DRs and DP SAA implementations, respectively. For infrastructure systems requiring large investments, this is a crucial improvement which can determine project feasibility or success. Interestingly, however, DRL did score worst in Value at Risk (VaR) and overall variance between scenarios (as seen by the shapes of the CDFs in

Figure 4.7) among all flexible alternatives. These indicators give some reason for concern and suggest further fine tuning of the solutions is needed to increase acceptability in a real-world setting, as VaR is often a very important metric in determining project investment, particularly for risk-averse decision-makers. The larger capacity expansion flexibility allowed by the DRL implementation may tend to favour looking for upsides at the cost of neglecting downsides, although ultimately the effect of increased upside potential remains dominant. Including flexibility strategies more focused on addressing downside risks, such as abandonment, could potentially aid in mitigating some of these effects and yield a narrower CDF profile.

Table 4: Multi-criteria design decision making assessment for the garage case study with stochastic demand over 2000 simulated scenarios.

Criterion	**Inflexible Baseline (1)**	**Flexibility DRs (2)**	**DP SAA (3)**	**DRL Flexibility Analysis (4)**	**Best Option**
ENPV	7.77	12.24	13.09	14.47	**4**
$\boldsymbol{VoF}$	0	4.47	5.32	6.70	**4**
VaR, 10%	-5.72	-0.21	0.28	-0.34	**3**
VaG, 90%	17.32	24.13	23.89	30.17	**4**

4.5 Understanding and Comparing DRL policies

This section presents some initial attempts at improving the interpretability of the DRL based solution for future practical implementation. It starts by looking at a few example stochastic demand scenarios and the respective capacity evolution strategies produced by Flexibility DRs approach. This analysis is then extended by retroactively evaluating the form of decision rules implemented by the DRL agent through a preliminary statistical analysis of their distribution. These outputs are then compared to the ones from the Flexibility DRs approach, providing some intuitive guidelines for operation and logic of the various solutions produced. Note that the DP SAA solution is not analysed in this section given it presents very limited interpretability (as discussed in 2.1.4) and generalizability for deviations from projected scenarios. Furthermore, based on Figure 4.7 and Table 4, it produces a very similar distribution of performance as the Flexibility DRs approach.

4.5.1 Capacity Evolution Scenarios

Capacity and demand evolution plots are used a starting point for evaluating the quality of flexibility strategies developed and comparing them to the inflexible baseline. This helps provide some visual intuition on the nature of the problem tackled, and differences in decision-making under uncertainty

among available methods. The example scenarios captured in Figure 4.8 suggests that both Designs 2 and 4 are well suited to adapt system configurations to realizations of uncertain demand over time. Note that designs here (and for the rest of this thesis) loosely refer to solutions obtained with the different computational design evaluation approaches. The Flexibility DRs approach, however, is more limited in the form of expansion decisions made. It contrasts with the DRL based solution relying on a complex NN parametrized policy, which based on progressive state observations dynamically selects the highest valued action. This leads to very limited demand-capacity mismatch throughout project life, although with a slight tendency toward overbuilding. In terms of economic performance, Design 4 produces a higher NPV for the high and low stochastic demand scenarios from Figure 4.8, albeit with a slightly lower one than Design 2 for the medium scenario. This suggests that the higher dynamicity in Design 4 is particularly well suited to generalize strategies in uncertainty realizations that are further from deterministic projections.

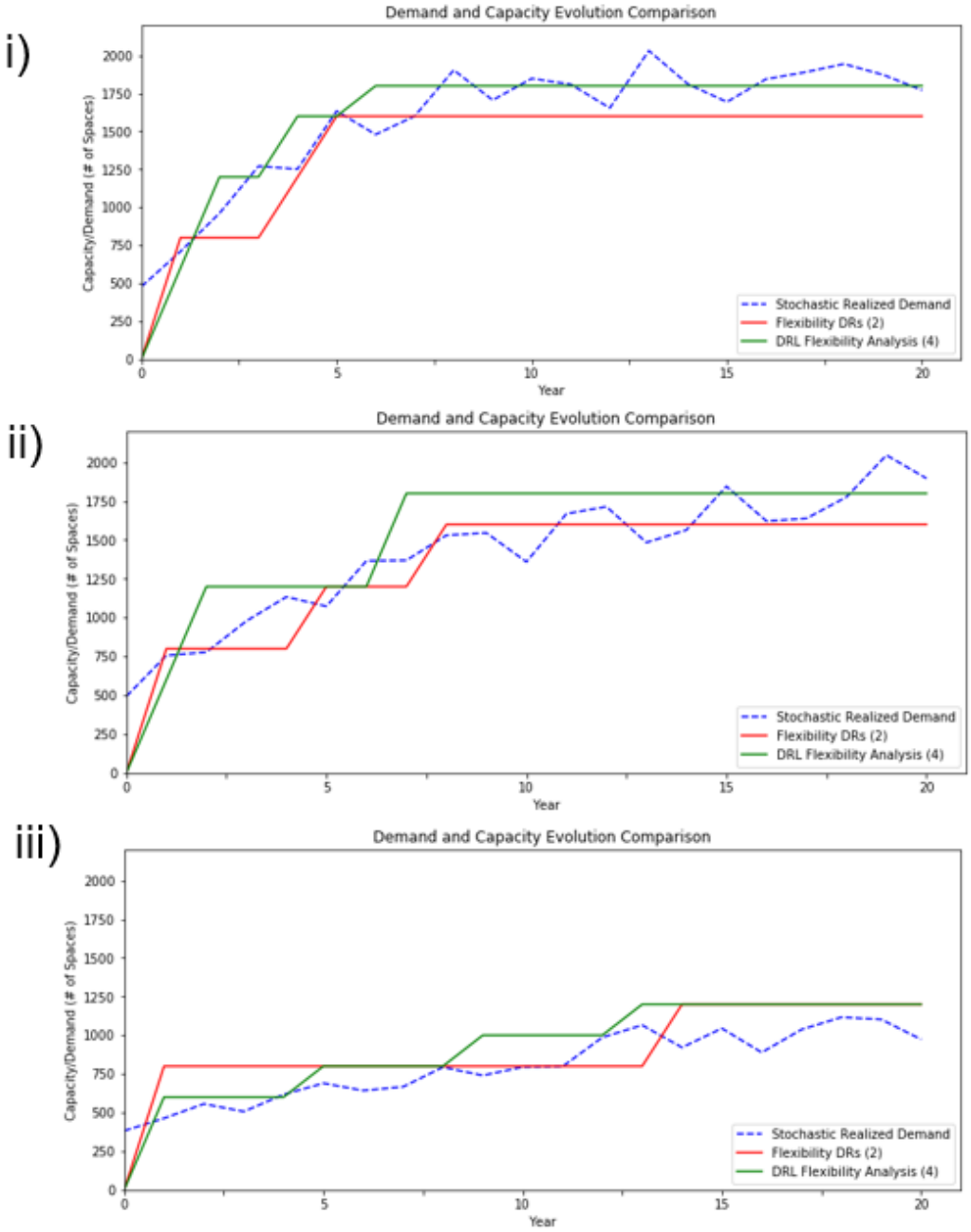


Figure 4.8: Example capacity evolution scenario using for the Flexibility DRs (red), and DRL Flexibility analysis (green) under different stochastic realizations of demand (blue), where i) captures a relatively high demand scenario, ii) a medium demand scenario, and iii) a lower scenario.

More specifically, Figure 4.8 i) captures a relatively *high* demand scenario which leads to early capacity deployment for both Designs 2 and 4, albeit through a more modular approach for the DRL solution. This ultimately results in a higher total capacity which closely matches realized demand past around year 7, capturing the upside potential for the remainder of system operation, albeit with slightly higher investment costs. Figure 4.8 ii), on the other hand, captures a *medium* demand scenario that is closer to deterministic projections. It highlights another important difference between the logic

of the two approaches, as Design 2 only exercises capacity expansion once there has been a sufficiently long period where it could not meet demand. Accordingly, it tends to be more focused on reducing downside exposure, while Design 4 deploys excess capacity which is not able to be filled in most years of system operation. Finally, Figure 4.8 iii) captures a relatively *low* demand scenario, where Design 2 deploys more capacity than Design 4 in early years, resulting in unnecessary high initial investment costs. The flexibility strategies employed are overall similar for both methods, with the same final capacity reached by year 15, and Design 4 presenting more modular deployment to discount costs into the future. Both approaches, however, yield solutions which are well adapted to this specific scenario, only presenting limited demand-capacity mismatch over system operation.

4.5.2 Decision Rule Distribution

As mentioned in the former sections, the DR approach detailed in 4.2.2 allows the most complementary evaluation of the performance of the DRL agent on the design problem. This is because in the real world, the most you could optimize for before the project is either the predefined capacity expansion plan or the DRs to determine capacity expansion strategy. Since DRs allow the decision maker to adapt the capacity expansion plan based on observations of actual realized uncertainties, a comparison with them is deemed most appropriate for methodological validation. The DRs implemented by Design 4 are retroactively approximated over 2000 stochastic out of sample scenarios to gain further insights on the DRL agent's decision-making process. The increased dynamicity of the proposed approach becomes more apparent through statistical analysis of simulated environment episodes, analysed through the set of DRs from Table 2, and implemented in Design 2.

Firstly, the DRL based solution is found to primarily deploy 2 or 3 levels in the zero stage decision based on initial demand, leading to an average rule of $\boldsymbol{f_0 = 2.63}$, significantly lower than the optimal starting capacity of 4 floors for Design 2. This is likely because the DRL action space provides more flexibility over how and when to build, thus the agent can quickly ramp up capacity depending on the reward signals and demand observed.

The number of years with demand greater than capacity before expansion produced by the DRL agent are plotted in Figure 4.9. Most often, demand exceeds capacity for two years before an expansion decision, resulting in an average rule of $\boldsymbol{d_r = 1.76}$. This is very close to the value found for Design 2, suggesting the two approaches produce comparable flexibility strategies, even though it should be noted that the DRL implementation does not require historical data to reach a similar conclusion. Consequently, Design 4 exploits a more complex internal policy network to determine expansion through the state observation in a particular year only. In nearly 40% of expansion time-steps, in fact, only 1 year of excess demand has been experienced. On the other hand, a wait time of 3 years only occurs in about 15% of timesteps exercising a flexibility strategy, perhaps based on the different magnitudes of expansion considered. Finally, the DRL agent appears more focused on capturing

upside potential, with no instances of 3 years with excess demand recorded other than if the maximum capacity constraint had been reached.

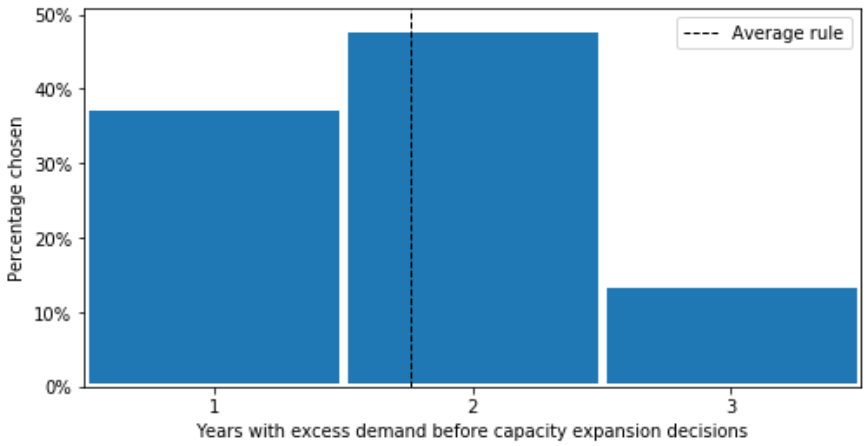


Figure 4.9: Histogram for d_r variable distribution of DRL agent policy estimated over 2000 scenarios.

The findings on the d_r variable distribution are likely correlated to the more modular action-space developed in 4.3.1, as compared to Design 2. This is further evaluated through Figure 4.10, showing the distribution of capacity expansion magnitude decisions over the same scenarios used in Figure 4.9. Note that, to provide a better comparison with the design vector from Table 2, decisions not to expand are excluded from this analysis, as they logically are the selected action in the majority of timesteps, significantly skewing the data distribution. Zero-stage decisions on starting number of floors are also omitted here as they were evaluated separately earlier in this section, although there is a potential correlation among the two variables. The results suggest that, most frequently, Design 4 expands capacity by only 1 floor, resulting in an average rule of $\boldsymbol{f_t = 1.67}$. Again, this is likely driven by the more dynamic nature of the DRL policy, as the agent is able to evolve system capacity more adaptably, not presenting the same decision constraints as Design 2. In fact, the DRL based policy is able deploy to the maximum capacity by year 6, if desired, even expanding just one floor per timestep. Expansion by 3 floors was seldom selected, which is somewhat intuitive as one would imagine there would more demand scenarios where it would be favourable to capitalize on economies of scale. Nonetheless, this may be simply due to the nature of discounting capital costs into the future. Furthermore, the high degree of dynamicity in DRL policies significantly reduces the need for such large capital investment commitments, even in very rapidly increasing demand profiles.

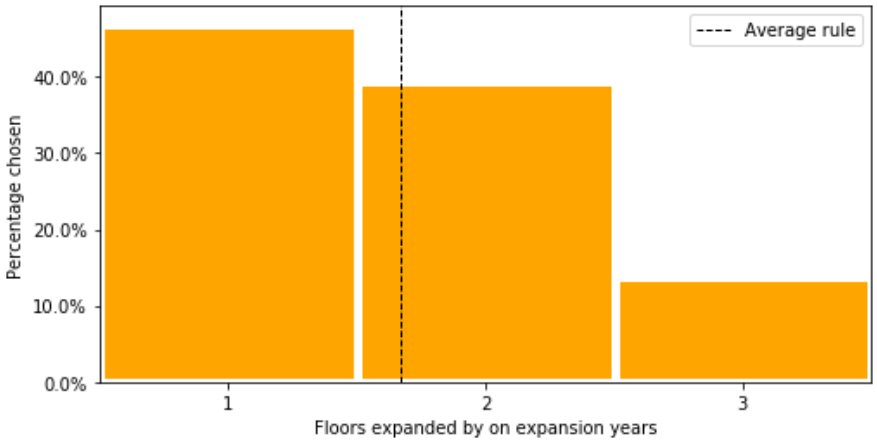


Figure 4.10: Histogram for f_t variable distribution of DRL policy estimated over 2000 scenarios.

This intuition is further supported by Figure 4.11, which shows the distribution of flexibility strategies for Designs 2 and 4 over the project operational life by estimating the expansion probability for each year, regardless of magnitude. The results suggest that a logical and intuitive policy is developed by the DRL agent (outperforming benchmarks from 4.4), with important similarities and contrasts to the Flexibility DRs approach. In essence, the distribution of expansion timings confirms the motivations from 1.5 that a NN parametrized policy (Design 4) can generate more dynamic decision making as opposed to restrictive and predefined DRs (Design 2). Flexibility options, it follows, are optimised over a broader range of system operational years and state observations. This suggests the development of a more generalizable policy which selects actions based on progressive uncertainty realizations. The graphical insights from Figure 4.11 are further decomposed below.

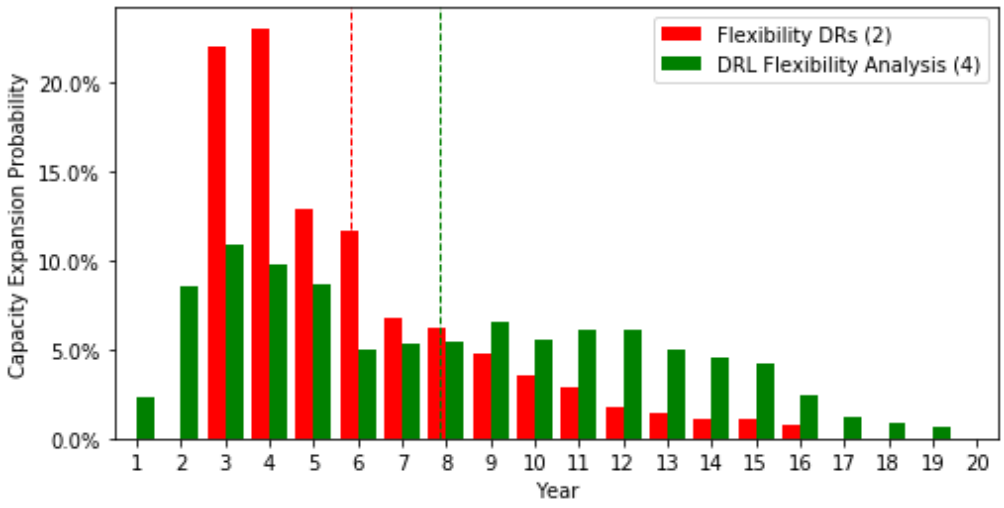


Figure 4.11: Distribution of expansion decisions over project life across 2000 simulated scenarios for Design 2 (red) and Design 4 (green). Vertical dashed lines are their respective mean expansion years.

More explicitly, capacity expansion is seldom pursued by Design 4 in the first year, driven by the stage-zero investment and an initial assessment of demand-capacity mismatch. Design 4, on the other hand, is always unable to expand in the first two years based on Procedure 1 from Figure 4.3. Accordingly, the highest expansion probabilities are found for years 2-5 (peaking at 3) and 3-6 (peaking at 4) for Designs 2 and 4, respectively. These are generally good timings for infrastructural investments, as they can generate profits over a long remaining project life, while exploiting uncertainty realizations and costs discounting. Past year 6, in fact, capacity expansion probabilities through the Flexibility DRs approach gradually decline to zero for years 17-20 (as constrained by Procedure 2 and Procedure 3 from Figure 4.3). On the contrary, capacity expansion probability is similar for years 6-12 for the DRL agent, and slightly higher for the latter ones as state observations capture stochastic demand evolutions. Past year 12 a similar gradual decline is observed as for Design 2, with the main difference being the minor probabilities recorded for years 17-20 for Design 4. This finding suggests that further fine tuning of the DRL agent policy may be necessary, as the capital structure of the project makes it unlikely to recoup investments made this late in its functional life.

4.6 Chapter Summary and Conclusions

This chapter presented the first implementation of DRL Flexibility analysis on a well-studied proof-of-concept infrastructure system, financially outperforming all benchmarked designs obtained through alternative state of the art approaches. Sustainability and Resilience metrics were not directly optimised or assessed for here, as the nature of the case study prioritized economic objectives from an investment perspective. Nonetheless, the general qualitative argument that flexibility can be a unifying paradigm for these concepts in engineering systems design , as discussed in 1.2, can be applied here too. On one hand, only expanding floors when they are needed helps improve the project sustainability, as emissions from construction are minimized compared to a fixed floor design. On the other hand, the resilience of the engineering system is also enhanced. Consider for instance a small earthquake happened that wiped out the top 2 floors of each garage, the flexible design would be able to restore capacity more quickly than the fixed one due to its embedded modular expansion ability.

At this stage, while the ability to enhance flexibility of the system by the DRL methodology is well supported, it is possible that it is overestimated, perhaps due to other issues at play. For example, the DRL agent may be overfitting to the environment's formulation. This was investigated by altering stochastic model parameters, and while the ENPV results varied under different assumptions, the DRL based solution consistently outperformed alternative approaches. The final results may have been due to some "luck" in the training process by finding high valued actions early on and may not be easily reproducible. The averaging of results from several separate training runs using different random initialization seeds, however, suggests this issue to be unlikely as final performance was similar across different agent trainings. The potential to scale the method up effectively to a real world highly complex design problem, where the action and state spaces may be much larger than that this

one, may remain a limiting factor or require substantial domain expert manipulation. This issue is addressed in more detail in Chapters 5-7. There may be logistical constraints which make the implementation of the DRL solution in real life more complex or more costly than it appears, as for instance there likely would be significant interruption in service if construction for expansion was to happen often, compared to all at once or in only two or three stages. The more risk-prone behaviour (i.e. overbuilding) experienced must be addressed to be suitable with risk-adverse decision-making preferences, which are often found in systems designers. Further, the fact that dispersion and variability increased for the DRL methodology compared to the benchmarked flexible designs brings must be addressed.

A notable advantage of the proposed methodology compared to the Flexibility DRs is that the form of the rules normally has to be pre-determined by system designers from a mix of creativity and intuition on the nature of the problem. In our proposed approach, however, all that is required is for designers to identify all possible decision-making actions, particularly those with a design implication, to then allow the agent to explore the environment and find the optimal ones over a much larger number of interactions. In this example problem, they are the decision to build 1-9 floors initially and then expand by 0-3 floors at any point during system operation. This can lead to more dynamic and adaptive policies under uncertainty, essential for the objectives of Flexibility in Design, while reducing the domain expertise required to formulate a flexible design and strategy. These contributions are particularly focused on complementing concept generation and design space exploration (Phases 3 and 4 in the proposed framework from 3.1).

The applicability and acceptability of computational insights into the design and decision-making process, nonetheless, remains an important point to address for its practical implementation. The DR distributions approximated in this chapter provide an avenue toward improved interpretability to non-experts, but the question of how they would actually be used during systems operation remains. One avenue would involve exposing the DRL agent to actual realized uncertainties and base operational strategies on its reactions. A more advanced statistical decomposition of the agent's actions under different conditions could also yield some valuable data-driven insights for visualization. Measuring and determining the acceptable level of risk tolerance for a project will be important in future work, and could be integrated into the reward function directly. Alternative NN architectures could be tested for different objectives. Policy gradient methods may also be explored to deal with higher dimensional environments. Accordingly, stochastic policies could be developed, as discussed in 2.3, which may be better suited to dealing with uncertainty for engineering systems design.

In light of the above, the proof of concept results presented in this chapter have demonstrated the potential of the proposed approach, justifying an expansion to a more complex case study to further explore its potential contributions.

Chapter 5

5 Methodological Expansion: Singapore Waste to Energy System

This chapter presents the expansion of the proposed methodology to the design and planning of a more complex engineering system under several sources of uncertainty, including the consideration of spatial factors in decision making. In short, it is motivated by the following question.

- RQ4: Does the proposed data-driven approach enhance expected value through flexibility even for more complex and higher dimensional engineering systems design problems?

This increased case study complexity as compared to Chapter 4 marks a shift in algorithmic approach taken, moving from value to policy based methods. The impact is an arguably more interpretable decision making procedure through the DRL agent, where the use of the underlying action probability distribution is explored as opposed to retroactive approximation of decision rules implemented. The impact of increasing action space size on training difficulty or tractability and final expected performance is also thoroughly evaluated here. The results are very promising on all aspects, confirming the outputs from Chapter 4, and allowing for a more complex evaluation of trade-offs among different design solutions. Finally, an initial attempt at out of sample policy testing is presented, drawing some important conclusions on generalizability of alternative methods. Note that this chapter is primarily based on the published peer-reviewed journal paper [10], accepted as part of a special issue on the role of artificial intelligence (AI) in engineering design. The full open source code to reproduce all results presented in this chapter can be accessed here[217].

5.1 Case Study Overview

This section presents the application of the proposed methodology to the analysis of a real-world WTE system design in Singapore. The case study analyzes tensions between two common approaches to systems design under uncertainty: 1) a centralized approach leveraging economies of scale, which is economically sound when EoS are strong under a set of predetermined scenarios, but requires significant capacity deployment early on, and; 2) a decentralized modular approach, which does not benefit as much from EoS, but helps managing uncertainty by deploying resources flexibly when needed, thereby extracting value from flexibility. The problem formulation is generic and applicable across a range of similar, distributed engineering systems, such as mini-grids, or other renewable energy systems. The benchmark system is a centralized large capacity WTE design located in the

western area of Singapore and processing food waste produced by the city-state. Food waste must be transported to the centralized location from each of six sectors comprised under the Public Waste Collection (PWC) scheme. The flexible systems are decentralized WTE designs that build a small-scale plant in each of the waste collection sectors. They differ in how the capacity is deployed over time and space, governed by either Flexible Decision Rules (FDR), or DRL dynamics. PWC workers can choose to transport the collected waste to a sector's Anaerobic Digestion (AD) plant rather than the main centralized one. Both types of systems receive revenues from the production of electricity and refuse collection, while the primary costs are associated with capacity expansion, waste transport, disposal, maintenance, and land rent. The decentralized designs allow a significant reduction in transportation and collection costs but see increased average cost per unit processed due to limited EoS. Their modularity, however, contributes to reducing exposure to downside conditions (e.g., lower amount of waste produced than expected in any given sector). As such, decentralized solutions have to coordinate and balance flexibility decisions across the six different waste collection sectors in the city to maximize ENPV over a 15-year period, with associated complex revenue and cost functions. Please refer to the original manuscript exploring this case study [208] for further details.

5.1.1 Model and Environment Formulation

The annual revenues AR^t associated with AD energy system operations are calculated according to (5.1) and are based on food waste collection R_D^t and electricity generation R_E^t sources of income. They are computed according to (5.2) and (5.3) respectively, where d_i^t is the recycled sector i food waste at year t and d^t the total recycled food waste in the city, as shown in (5.4). The central site capacity is given by θ. The food waste collection income is determined from tipping fee P_{to} and additional revenues are found from the electricity generation rate E_g and the unit selling price of electricity P_e.

$$AR^t = R_D^t + R_E^t \tag{5.1}$$

$$R_D^t = d^t P_{to} \tag{5.2}$$

$$R_E^t = \min(d^t, \theta) E_g P_e \tag{5.3}$$

$$d^t = \sum_{i=1}^{6} d_i^t \tag{5.4}$$

The annual system cost AC^t shown in (5.5) is found from summation of the transportation cost C_{Tran}^t, disposal cost for waste residue C_D^t, the land cost C_L^t, and operational and maintenance (O&M) cost C_{OM}^t. The break down how each of these components is calculated is given by (5.6)-(5.11).

$$AC^t = C_{Tran}^t + C_D^t + C_L^t + C_{OM}^t \tag{5.5}$$

$$C^t_{Tran_{centralised}} = C_{fuel} \sum_{i=1}^{6} \frac{d_i^t}{Cap_v} (D_{co_i} + D_{Tr_i}) \tag{5.6}$$

$$C^t_{Tran_{DecentralizedFlex}} = C_{fuel} \sum_{i=1}^{6} \frac{d_i^t}{Cap_v} D_{co_i} + C_{fuel} \sum_{i=1}^{6} \frac{d_i^t - \theta_i^t}{Cap_v} D_{tr_i} \tag{5.7}$$

$$C_D^t = (min(d^t, \theta)\,(1 - \omega + \varepsilon) + max\,(d^t - \theta, 0))C_{dis} \tag{5.8}$$

$$C_L^t = \rho\theta \tag{5.9}$$

$$C_{OM}^t = \varphi C^0(\theta) = \varphi[K(\theta)^\alpha] \tag{5.10}$$

$$\theta_i = \theta/6 \qquad i=1, 2,\cdots, 6 \tag{5.11}$$

For the centralized design, total transportation cost is computed from the sum of waste collection cost within each site, as well as transporting waste from each site to the central plant according to (5.6). D_{co_i} is intra sector i collection distance, and D_{Tr_i} is the distance for collection from sector i to the main site. In the decentralized system, transportation costs are found by first estimating excess untreated waste in each sector i as $(d_i^t - \theta_i^t)$, and resulting number of trips required to transport the waste volume $\frac{(d_i^t - \theta_i^t)}{Cap_v}$ to yield total collection truck travel distance, as shown by (5.7). C_{fuel} is the unit cost for fuel consumption. The cost of disposal C_D^t is determined as a function of the purity rate ω, residue rate ε and the unprocessed waste due to system wide capacity shortages for year t from (5.8). All waste that must be disposed of is assumed to be incinerated at unit cost C_{dis}. Land rental costs C_L^t in (5.9) are calculated based on installed capacity levels θ at year t, and the unit land rental fee ρ. Finally, operational costs are found through (5.10) as a function of total installed capacity accounting for EoS and based on a constant factor φ, or the sum of the ones in each sector, as given in (5.11).

Recycled food waste (FW) – also referred as the *demand* for food waste processing (d^t)– is simulated using the standard GBM approach introduced in 3.4.4 , representing the overall increasing trend in food waste production, while accounting for random shocks and uncertainty. Historical data is used for building the stochastic model and enables direct comparison with results from the original case study. The mean growth rate μ and volatility σ are estimated from previous food waste generation patterns in Singapore (μ = 12.3%, σ = 16.3%) using standard statistical regression. The random variable dW_t is sampled from a standard normal distribution here. ξ^0 (Or d^0) is the total recycled food waste as of 2013 in Singapore ($\xi^0 = d^0 = 274$ tonnes per day, or tpd), which is also the assumed starting point for the study. Using these assumptions, 2000 i.i.d scenarios are generated over the 15-year time horizon of the project, showing evolution of recycled food waste in each sector. The

high degree of stochasticity associated with this engineering systems design problem is graphically captured by Figure 5.1. It shows how demand for FW processing can vary significantly across different sectors in a particular scenario (a), as well as the potential large variability in total city-wide recycled FW under uncertainty (b). This suggests that decentralized solutions will need to account for spatial factors to maximize expected value, in addition to overall trends across the system evaluated.

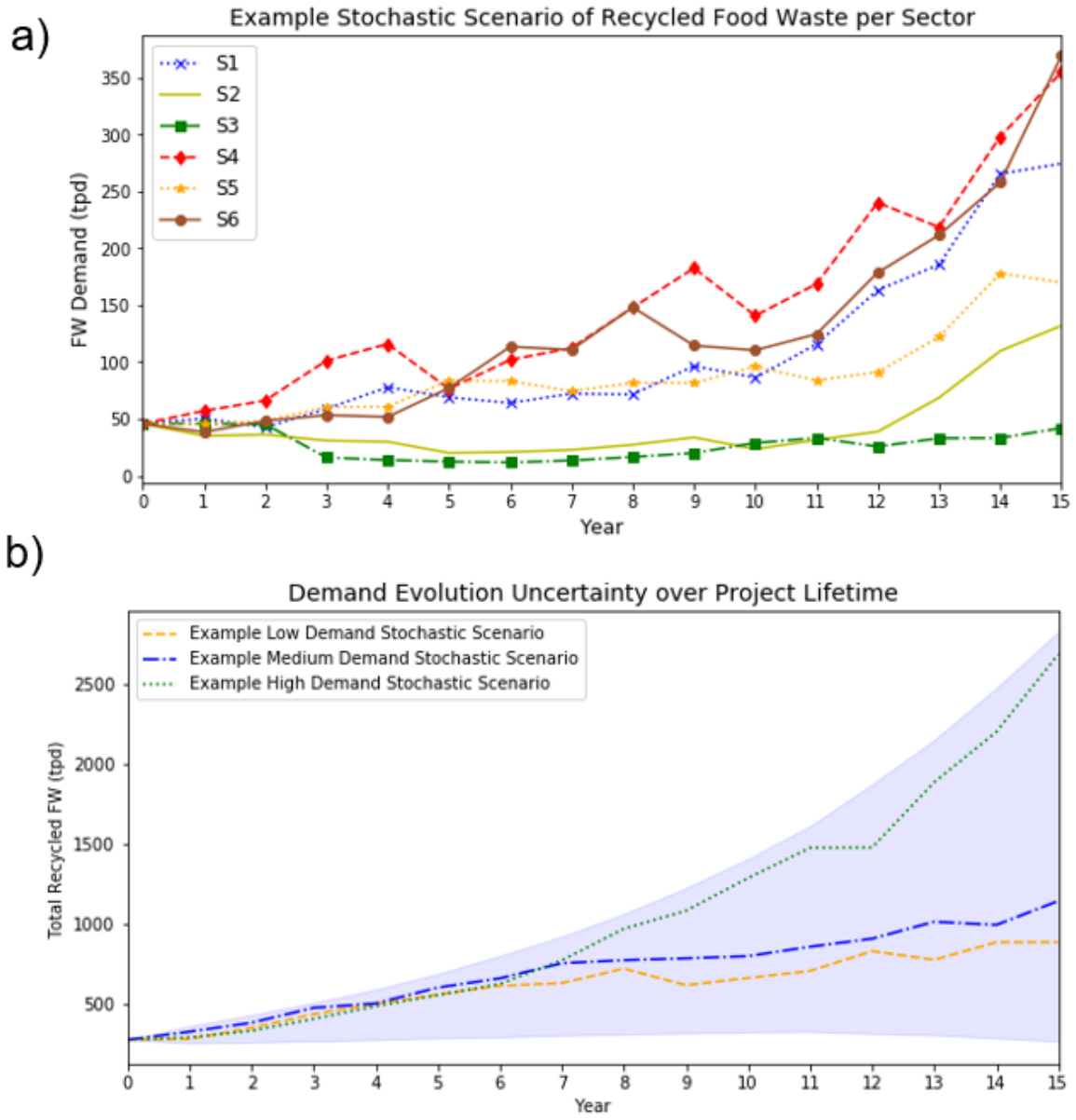


Figure 5.1: (a) Example stochastic scenario of FW processing demand by sector ; (b) city-wide stochastic evolutions over 2000 simulated scenarios including low, medium, and high example scenarios, where blue shaded area represents 95% confidence interval.

Please refer to 11.3 for the complete list of assumptions and parameters used for the model detailed in this section.

5.2 Benchmarking

The quality of solutions obtained with the proposed methodology is obtained by benchmarking against the design alternatives listed below.

A. Stochastically optimal centralised inflexible design.
B. Stochastically optimal decentralised design with Flexibility DRs (FDR).

No benchmarking implementation relying only on dynamic programming is presented for this case study, given the significant dimensionality of the problem – which is also an important limitation of such method. Additionally, it was shown previously that a FDR approach can estimate to a very similar degree the value of flexibility obtained using dynamic programming [62], thus removing justifications for such comparisons past the proof of concept methodological validation presented in Chapter 4. Additionally, we note that comparison with a centralized flexible and decentralized inflexible solution is also available in the original work by Cardin and Hu [208]. These are not presented here to focus on the most relevant alternative methodologies to the objectives of this thesis – as a complete analysis of the trade-offs introduced by EoS is outside the scope of this work.

5.2.1 Centralized benchmark design

The objective function for capacity planning model of the centralized design with no expansion flexibility under uncertainty is given in (5.12), based on the model detailed in 5.1.1. This optimised using the DE algorithm, as in Chapter 4, yielding the same solution as the original study – that is a total capacity $\boldsymbol{\theta = 600\ tpd}$ installed in Sector 1, with no changes over the 15-year project life evaluated. Note that this outperforms decentralized inflexible solutions under uncertainty due to EoS effects, as detailed in Cardin and Hu [208].

$$\max ENPV = \frac{1}{S}\sum_{s=1}^{2000}\left(\mathcal{C}^0(\theta) + \sum_{t=1}^{T}\frac{AR_s^t\left(\theta_s^t, \xi_s^t\right) - AC_s^t\left(\theta_s^t, \xi_s^t\right)}{(1+\lambda)^t}\right) \tag{5.12}$$

5.2.2 Flexibility decision rules

Based on the results from the earlier section, the potential of the proposed methodology on this higher complexity case study is best assessed by comparing with outputs from the FDR approach. The variables considered and the values analysed for optimisation are summarised by Table 5 below.

Table 5: Decision rules variables and their ranges considered for the WTE case study.

Decision Rule Variable	Explanation	Levels Evaluated
$\boldsymbol{\theta_1}$	Starting capacity in main site (i.e., Sector 1)	200-600
$\boldsymbol{\beta}$	Total FW processing demand and capacity mismatch	0-3
$\boldsymbol{\tau}$	Threshold for decentralized expansion in modules	0-5
$\boldsymbol{\gamma}$	Expansion magnitude in number of modules	1-5

The objective function using this approach is formulated according to (5.13), which is optimised over the 2000 stochastic scenarios (S) generated. The remaining variables are defined in 5.1.1, calculating the discounted ENPV as the difference between annual revenues and costs less the initial investment.

$$\text{Max } ENPV = \frac{1}{S}\sum_{s=1}^{2000} -(C^0(\theta_{1,s}) + \sum_{i=1}^{6}\sum_{t=1}^{T} \frac{AR_{i,s}^t(\theta_{i,s}^t, \xi_{i,s}^t) - AC_{i,s}^t(\theta_{i,s}^t, \xi_{i,s}^t) - C_{\text{exp},i,s}^t}{(1+\lambda)^t}) \quad (5.13)$$

The logical progression of the approach on this case study is graphically captured in Figure 5.2. **Procedure 1** determines first whether there is a need to expand capacity at a system level based on overall demand and installed capacity. **Procedure 2** is used to decide whether the expansion should proceed in a decentralized manner, or at the main site only (i.e., sector 1). If a decentralized expansion is chosen, **Procedure 3** determines in which sector to add capacity, up to the maximum allowed θ_{max}. One decentralised module has a capacity of 50 tpd for the calculations below. Note that the above rule contains a nested if statement, so that **if** *there is a capacity mismatch at the system level, but no sector satisfies the required capacity mismatch*, **then** *capacity is deployed at the main site (sector 1)*, **else** *no additional capacity is added to the system.*

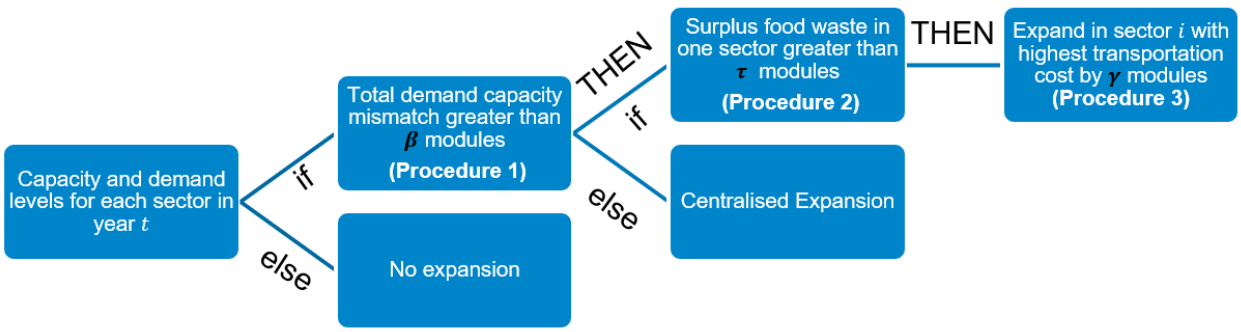


Figure 5.2: Visualization of FDR implementation to determine expansion decisions on WTE case study. DR variables are highlighted in black for improved graphical clarity.

The optimal values for the variables detailed in Table 5 are given in (5.14), and their implication will be discussed in more detail in 5.4, although they provide readily usable guidelines for decision makers in terms of both initial design and triggering signals to exercise flexibility during system operation. .

$$[\theta_1, \beta, \tau, \gamma]^* = [\,200, 1, 4, 0.5] \tag{5.14}$$

5.3 DRL Implementation

5.3.1 DRL Flexibility Analysis Formulation

As discussed in 3.2, the Open AI gym format is used to develop the environments representative of the engineering systems design case study tackled in this section. The steps for the proposed DRL flexibility analysis methodology are further detailed below.

1. **Environment Formulation:**
    i. State Space: The observation space which the DRL agent exploits for decision making under uncertainty is formulated through a box observation space, representing the installed capacity ${\theta_i}^t$, and realized stochastic demand ${\xi_i}^t$ in sector i for year t, respectively, as for the benchmarked approaches. The capacities within the observation are normalized by the minimum (0) and maximum (600 tpd) values for each site, while the upper bound on demand for all sectors is found from simulating 10,000 scenarios (independent of the environment) and finding the highest demand value in each sector to normalize those state variables. Time awareness for the agent $T - t$ is also implemented to address convergence stability and produce better value approximations for near terminal states, maintaining the Markov property. The complete state variable vector s^t is thus defined according to (5.15) below.

$$s^t = [{\theta_1}^t\,,\, {\xi_1}^t\,, {\theta_2}^t\,,\, {\xi_2}^t, {\theta_3}^t\,,\, {\xi_3}^t, {\theta_4}^t\,,\, {\xi_4}^t, {\theta_5}^t\,,\, {\xi_5}^t, {\theta_6}^t\,,\, {\xi_6}^t, T - t] \tag{5.15}$$

    ii. Reward Function: In order to reflect decision-makers objectives in this context, the reward function r^t for each time step is defined as shown in (5.16), leading to an ENPV maximization problem. The equivalent discount rate as for benchmarked approaches is implemented here. Furthermore, to address agent building past the maximum allowable capacity to capture potential upsides, an imaginary reward penalty of –S$1 is given in time steps where the capacity in a specific sector is at θ_{max}, and the selected action is not equal to 0. The implications and effectiveness of this approach on the case study tackled in this chapter are further discussed in 5.5.

$$r^t = \begin{cases} AR_i^t(\theta_i^t, \xi_i^t) - AC_i^t(\theta_i^t, \xi_i^t) - C_{exp_i}^t & if\ \theta^t{}_i \leq \theta^{max}{}_i \\ -S\$\,1 & otherwise \end{cases} \quad (5.16)$$

2. **Uncertainty Quantification:** The characterization and simulation of uncertainty on this problem is estimated as detailed in 5.1. The assumptions presented in Table 14 from 11.3 are thus used to generate stochastic scenarios according to the GBM modelling approach detailed in (3.11)-(3.12), leading to the same evolutions as graphically captured in Figure 5.1. One important difference, however, is that a much larger number of scenarios can be tractably used in the policy optimisation process – meaning uncertainty realizations are not restricted to the same 2000 i.i.d. scenarios as used for the benchmarked approaches.
3. **Action Space Creation:** In light of the methodological expansion objectives of this chapter, two separate action spaces (and respective environments) are developed to enhance flexibility in this engineering systems design. The reasoning is further discussed below.
    i. DRL-LF: An initial model referred as DRL-LF (i.e., low flexibility) is developed to ensure performance is benchmarked against standard methods appropriately, where capacity expansion decisions are restricted to the same magnitude as for the FDR model (i.e., 200 tpd per expansion). The reasoning is to evaluate how dynamic the agent's decision-making process is when limited to the same actions as for the FDR model, to see what insights may be gained simply on optimal exercise time. The starting capacity decision is left up to the agent in this initial implementation, although the magnitude is kept equal to that found with the FDR model, with the decision based on which sector to construct at $t = 0$. The discrete action set for the DRL-LF agent is, therefore, defined according to (5.17), with 7 potential decisions to select from in any given timestep.

$$a^t = \begin{cases} 0 \rightarrow & No\ Expansion \\ I \rightarrow & Expand\ by\ 200\ tpd\ in\ sector\ i \end{cases} \quad \text{for t} \in \text{T} \quad (5.17)$$

    ii. DRL-HF: Building upon this initial implementation, a more complex DRL model is introduced allowing for greater flexibility (i.e., DRL-HF for high flexibility). The added complexity stems from the fact that the magnitude of capacity expansion decisions is not preset at 200 tpd. Rather, the agent can explore several different expansion levels. This adds much more dynamic decision making, subsequent systems design implications and can present a more holistic and complete overview of the potential added value for this methodology. In this environment, the observation space is kept the same, but the action space is changed to multi-discrete decisions such that $I_j = [50, 100, 150, 200]$ respectively for each sector and only one action allowed to be selected in each time step, as captured by (5.18). The

capacity increments are in discrete values of 50 tpd based on the AD technology considered (see 5.1). It should be noted that for DRL-HF, the original constraint where max capacity had to be limited to that of the benchmark design was relaxed to evaluate more decision-making possibilities. A constraint on max sector capacity of 600 tpd per sector is still implemented to ensure benchmarking with a comparable system. This means, however, that the results should be taken more as a measure of a relative untapped potential for flexibility in the system given the original constraints were significantly relaxed.

$$a^t = \begin{cases} 0 \rightarrow & No\ Expansion \\ I_j & \rightarrow Expand\ by\ j\ tpd\ in\ sector\ i \end{cases} \quad \text{for t} \in \text{T} \tag{5.18}$$

5.3.2 Algorithmic Approach

In this chapter, we transition to a policy gradient (PG) method based on the high dimensionality in action (7 and 19 possible actions for DRL-LF and DRL-HF models, respectively) and observation spaces (13 state variables). Value based methods typically struggle in this setting as the number of potential state-action pairs grows exponentially with action and observation space size, creating very large computational and memory requirements. While policy gradient methods may be less sample efficient, thus normally requiring more simulated episodes to learn a usable policy, this is not an issue here since design simulations are relatively inexpensive for this model. In a highly stochastic environment such as here, PG methods can be advantageous as they output probability distributions over different actions to represent uncertain agent dynamics via a stochastic policy. Initially, this stochastic policy presents a high degree of randomness to allow agent exploration of different potential states and build a more complete representation of the environment dynamics. Over the course of training, this distribution converges to a deterministic policy and suggested action for each state, as the agent exploits rewards it has already experienced based on the policy update rules defined. This results in better convergence stability as well as a more generalizable policy than obtained via value methods. This class of DRL methods also helps addressing perceptual aliasing, whereby it can better differentiate between very similar states to yield the optimal action probability distribution [69].

PG methods, however, do present a few technical issues. Excessive policy changes can hinder training, as it may be difficult to map changes between policy and parameter spaces directly. Even very small differences in parameters can cause significant variations in performance, and a single overly large “incorrect step” for the policy update can thus lead to great errors in approximated action values. Improper learning rate may anneal or overinflate the gradient, which may lead to policies being stuck in local optima. Trust Region Policy Optimization (TRPO) is an iterative method

developed to address some of these limitations [149]. It uses a combination of the Minorize-Maximization (MM) algorithm and a "trust region" for policy updates. The MM algorithm approximates a lower bound for the received reward to guarantee monotonic improvements in policy, such that with each iteration it either improves or stays the same, theoretically converging to the optimal policy. This lower bound, also known as surrogate advantage, is calculated using (5.19):

$$\mathcal{L}\ (\varphi_k, \varphi\ ) = \frac{1}{1-\gamma}\ E_{s\ ,a\sim\ \pi_{\varphi_k}}\left[\frac{\pi_\varphi\ (a|s)}{\pi_{\varphi_k}(a|s)}\ A^{\pi_\varphi}(s,a)\right] \tag{5.19}$$

Here π_φ denotes the updated policy and π_{φ_k} the old policy used to estimate improvements. Generalized Advantage Estimation (GAE) is an improvement on these estimators that uses an exponentially weighted average of the *k*-step estimators [150], as captured by (5.20).

$$\widehat{A_t}^{GAE(\gamma,\lambda)}\ = (1-\lambda)\left(\widehat{A_t}^1 + \lambda\widehat{A_t^2} +\ \lambda^2\widehat{A_t^3} + \cdots\right) = \sum_{t=0}^{\infty}(\gamma\ ,\lambda)^l\delta_{t+1}^V \tag{5.20}$$

The trust region adjusts the step size for policy updates based on the objective function curvature and differences with the approximated one. This ensures the magnitude of policy updates is proportional to the confidence level within the trusted region. Anything outside this trusted region, where the approximation error is deemed too large, fails the improvement guarantee, and thus is not considered. To quantify the difference among the old and new policies for states already visited, restricting the size of the update step based on the trust region, the Kullback–Leibler (KL) divergence is calculated according to (5.21):

$$D_{KL}(\varphi||\varphi_k) =\ E_{s\sim\pi_{\varphi_k}}[D_{KL}(\pi_\varphi(\cdot\ |s)|\left|\pi_{\varphi_k}(\cdot\ |s)\right)] \tag{5.21}$$

Combining these equations yields the theoretical TRPO update given in (5.22) as:

$$\varphi_{k+1} = arg\,max_\varphi\ \mathcal{L}\ (\varphi_k, \varphi\ ) \tag{5.22}$$

$$\text{s.t. } D_{KL}(\varphi||\varphi_k) \leq \delta$$

where δ represents the KL-divergence limit. TRPO introduces a backtracking line search to the update rule, which addresses approximation errors introduced via Taylor series expansion required from Lagrangian duality theory [149].

The full pseudocode for the TRPO implementation is shown in Algorithm 2 below.

Algorithm 2: Trust Region Policy Optimization (TRPO) with generalized advantage estimation (GAE) applied to WTE case study, adapted from [149].

Initialize: Starting policy parameters φ_0, value function parameters $\emptyset_0$, KL divergence limit δ, backtracking coefficient α, maximum number backtracking steps K

for k = 0, 1, 2, … **do**

  Generate stochastic demand scenarios ${\xi_k}^i$ according to (3.11)-(3.12)

  Collect set of trajectories D_k by running policy π_k (φ_k) on the environment including decision for starting capacities in each sector at $t = 0$

  Estimate advantage ${A_t}^{\pi_k}$ using (5.20) based on current value function $V_{\emptyset_k}$

  Formulate estimates for policy gradient $\widehat{g_k}$ average KL-divergence Hessian-vector $\widehat{H_k}$ product function based on current sample

  Use conjugate gradient algorithm to compute divergence gradient and trust region

  Compute proposed policy change step Δ_k

  Update the policy by backtracking line search as $\varphi_{k+1} = \varphi_k + \Delta_k$

  Refit value function based on observed trajectories and policy

end for

Figure 5.3 gives an overview of the logical progression of the proposed approach following Algorithm 2, for both the DRL-LF and DRL-HF models. It compares and contrasts with the logic depicted for the FDR model in Figure 5.2. The state observation in any given year becomes an input into the feedforward NN parametrized policy, where an action is sampled from the underling probability distribution. One additional hidden layer is implemented in DRL-HF compared to DRL-LF to account for added action-state complexity, with output layers matching action space size. Some important differences should also be noted in respect to the value based DRL-method implemented in 4.3 and depicted by Figure 4.4. Firstly, agent-environment interactions served for value function updates in the former, while the policy is directly related updated in this chapter. Secondly, rather than simply returning a selected action, the full decision-making probability distribution returned by TRPO can be used to gain further computational insights for real world implementation, as further discussed in 5.5.

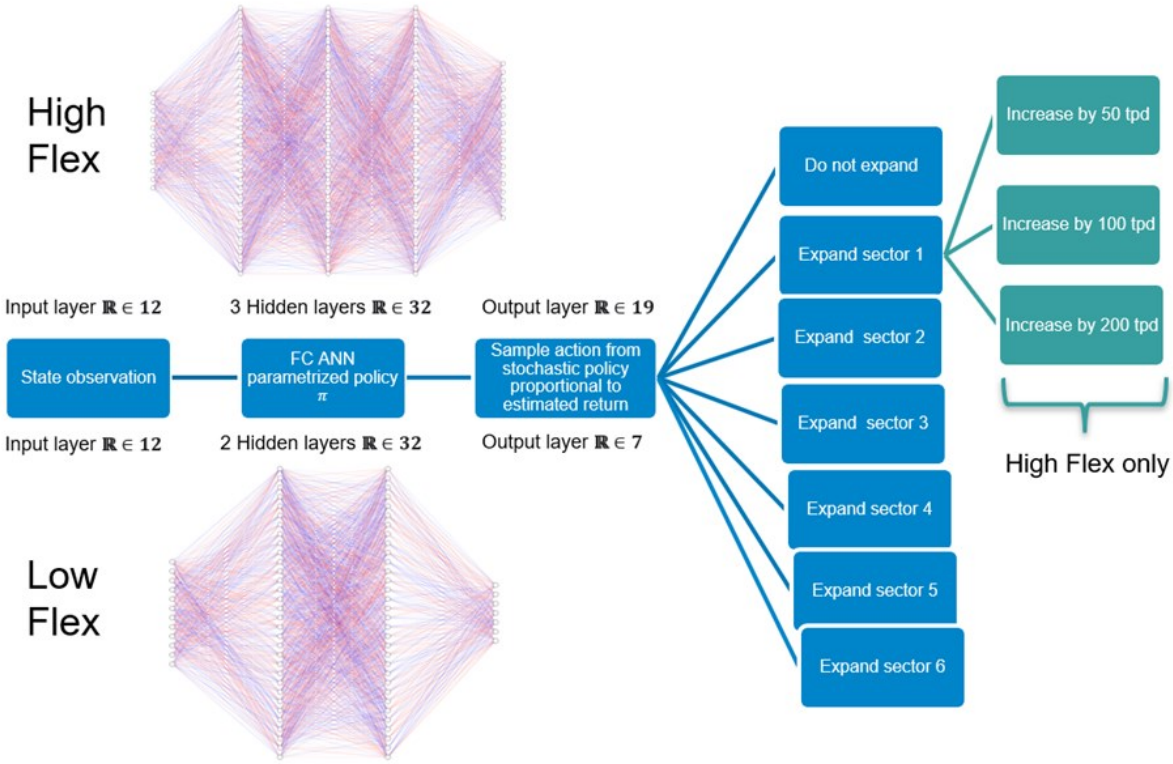


Figure 5.3: DRL decision process for both the low flex (DRL-LF) and high flex (DRL-HF) implementations. Magnitude of expansion limited to 200tpd in blue boxes for low flex implementation, no other expansion levels examined.

5.3.3 Training and Testing

Given the objective to yield a generalizable policy for agent-environment interactions, i.i.d. stochastic scenarios for FW processing demand are generated, as seen in line 3 of Algorithm 2. This leads to an infinite number of potential state action pairs, further justifying the use of a PG approach. DRL agent training is executed using $\boldsymbol{\delta} = 0.01$, $\boldsymbol{\alpha} = 0.8$ and maximum number of backtracking steps $\boldsymbol{K} = 10$. Remaining hyperparameters (please refer to 11.4 for a complete list) are found using Optuna looking at 100 different experiments for 100,000 training timesteps to systematically determine the best performing ones in this case[170]. Empirical results suggest this to be a much more effective approach than the grid search implemented in Chapter 4. The training process is fully non-anticipative, in that there is no look ahead bias for the agent, and operational decisions are purely based on state observations at the that time step. The policy built inside the NN, therefore, while learning from past mistakes and interactions, is not based on any information on future states, as the i.i.d. scenario generation method ensures no trajectory inside the environment is the same. All NN parameters θ are updated using the Adam optimizer[124].

Computational budgets of 1 and 1.5 Million timesteps are implemented for DRL-LF and DRL-HF agents training, respectively. As performed on a standard laptop Intel(R) Core(TM) i7-8665U CPU @

1.90GHz, the complete training process takes approximately 2 hours for the DRL-LF agent and 3 hours for the DRL-HF agent. The learning curves for the first 1 million steps for both agents are shown in Figure 5.4, averaging results from 5 separate training runs with different random initialization seeds and same hyperparameters. Note that only the best training run is used for relative performance analysis in 5.4. The graphic suggests the training process for both agents is successful, although with a few crucial differences. Firstly, the DRL-HF agent exhibits a more negative initial reward evolution, as the policy optimisation process needs to explore a higher dimensional action space and thus is more likely to incur negative episode NPVs. In fact, the policy for the DRL-LF agent seems to converge past around 250,000 timesteps, much sooner than for the DRL-HD agent. The DRL-HF agent starts outperforming the DRL-LF agent after around 750,000 timesteps of learning, and only begins converging in the final third of its training process (i.e. from 1 million to 1.5 million timesteps- not pictured here for visual clarity). Finally, it should be noted that the higher dimensionality of the DRL-HF agent results in an overall increased variance among project NPVs in different simulated episodes, as further detailed in the following section. The increased complexity of training the DRL-HF agent as opposed to DRL-LF can also be attributed to larger NN developed to capture its decision making policy (and corresponding number of parameters), as pictured in Figure 5.3. These larger networks and dimensionality of the problem tacked, it follows, motivated a higher computational budget for DRL-agent training as opposed to Chapter 4.

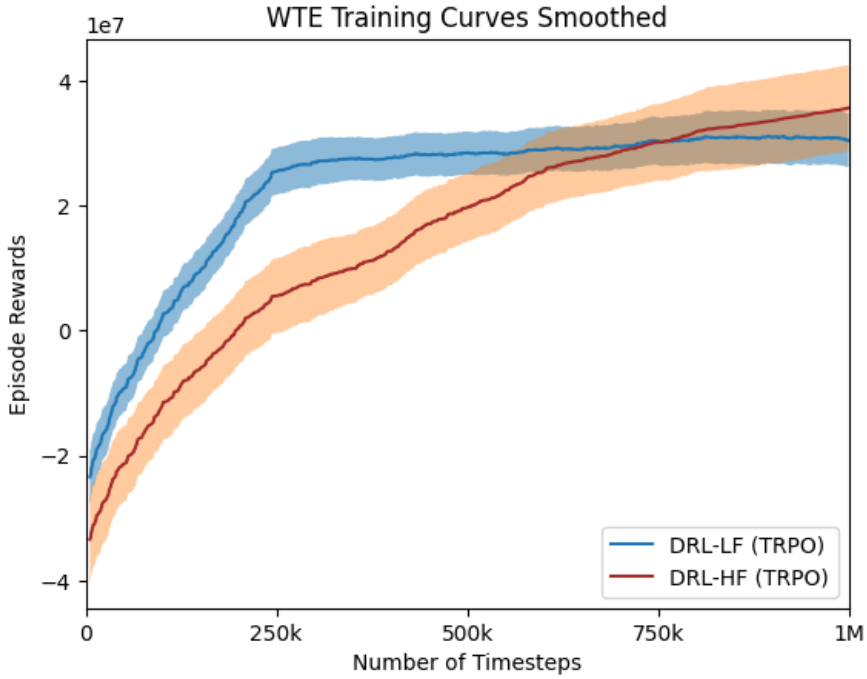


Figure 5.4: Training reward evolution for first 1 million steps of DRL-LF and DRL-HF agents on WTE case study. The solid lines show the smoothed averages using a window size of 100 episodes. The shaded areas show the standard deviations among different training runs.

The reward penalty for constraint violation captured by (5.16) is removed during the testing phase once the DRL agent has learned the form of the constraints implicitly throughout the training process. As such, the true ENPV obtained through the proposed approaches is evaluated for benchmarking against the results from Cardin and Hu. Additionally, the same 2000 i.i.d. scenarios are used for a more comparative evaluation. To do this, the agent-environment interaction framework is slightly modified as successive state observations are explicitly defined for the $\xi_i{}^t$ parameters in the observation array given in (5.15) , while the $\theta_i{}^t$ parameters are determined based on agent actions and expansion decisions. The different time-resolutions implemented for the case study presented in this chapter, as well as how they relate to each other, are summarized by Figure 5.5.

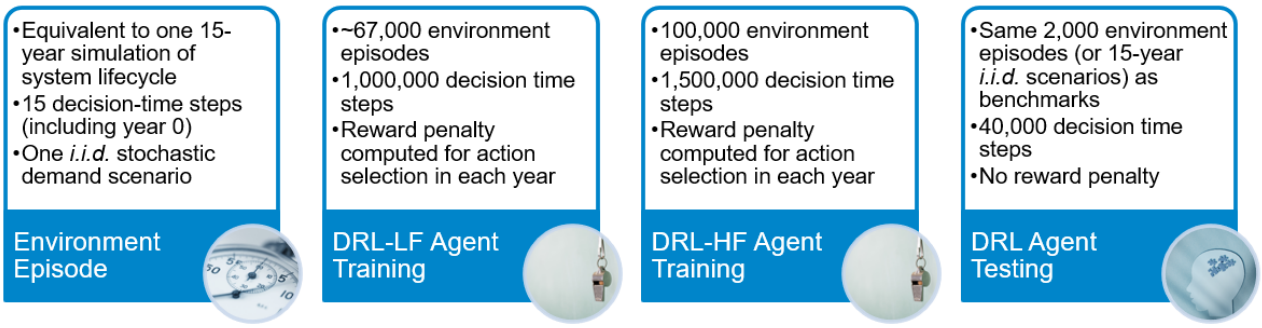


Figure 5.5: Summary of timestep resolutions from DRL agent training and testing on WTE case study.

5.4 Relative Performance Analysis

5.4.1 Summary of Results

The combined performance for the best Centralized Inflexible design, best performing Decentralized FDR design, as well DRL-LF and DRL-HF models proposed in this study are summarized in Figure 5.6. The simulation outputs confirm that embedding capacity expansion flexibility improves system performance significantly under uncertainty, with all flexible designs improving ENPV compared to the benchmark system. Implementing the DRL based flexibility design process yields even greater performance improvements and extracts further value from flexibility than obtained with the FDR model.

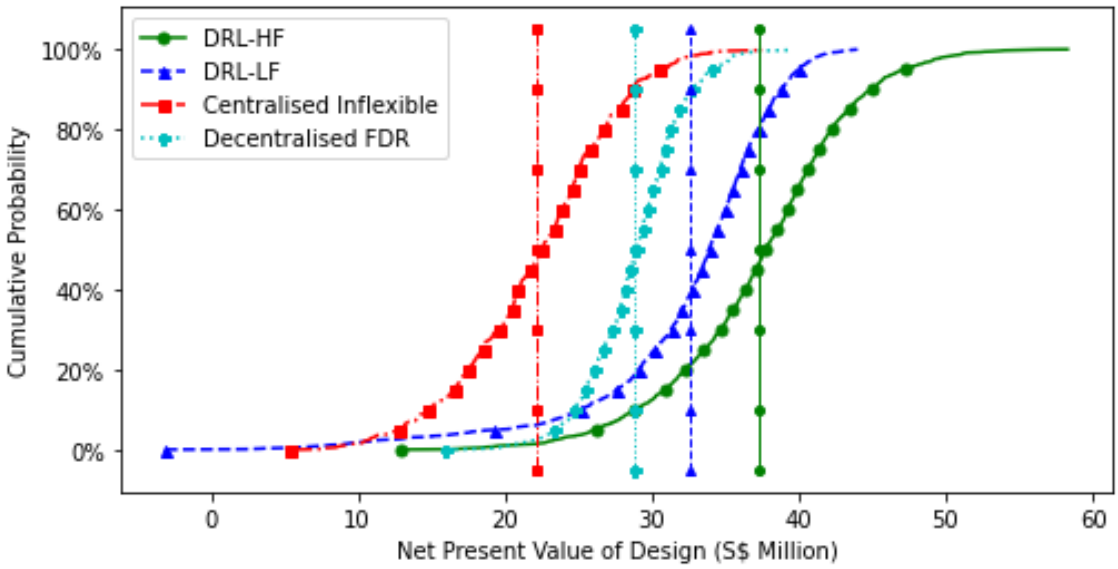


Figure 5.6:Cumulative distribution functions showing lifetime performance of different design alternatives over 2000 simulations. Vertical lines with corresponding markers represent ENPV values.

Table 6 provides further details on the stochastic performance results along different risk utility metrics. This is important to account for the possibility that decision-makers may show different risk preferences. For risk-neutral decision-makers, the analysis shows that both DRL models provide improved ENPV and extract additional VoF from uncertainty as compared to previous solutions. All flexible models improve exposure to downside risks (Value at Risk - VaR, 5%) as compared to the centralized system, which is important for risk-averse decision-makers. They also help risk-seeking system operators capture better upside potential (Value at Gain - VaG, 95%). Unsurprisingly, the DRL-HF model is able to extract the most value from uncertainty, along each metric. The results are discussed further below.

Table 6: Design decision making table for WTE system (all values in Million S$)

B	**(1) Centralized Inflexible**	**(2) Decentralized FDR**	**(3) DRL-LF**	**(4) DRL-HF**	**Best Solution?**
ENPV	22.19	28.73	32.55	37.45	DRL-HF
VoF	0	6.54	10.36	15.26	DRL-HF
VaR, 5%	13.27	23.18	19.33	26.59	DRL-HF
VaG, 95%	30.41	34.25	40.06	47.11	DRL-HF

The numerical results for the two DRL systems further confirm that the DRL methodology has the potential to complement and improve existing approaches to design complex engineering systems for flexibility. Looking back at the objectives from the beginning of this chapter, the analysis shows that the proposed approach does help support design of complex real-world systems operating under uncertainty. The ENPV compared to the Centralized Inflexible benchmark is improved by **47% and 69%** for the DRL-LF and DRL-HF implementations, respectively, and by **13% and 30%** respectively as compared to the Decentralized FDR system.

5.4.2 Out of Sample Testing

Comparing the original results in the study by [208] to the DRL solutions, the performance improvements are visually striking, but their interpretation should be conducted carefully. Among others, the stochastic nature of the TRPO policy means that there are cases where the DRL agent violates the constraints imposed on maximum capacity. This is a result of the distribution for action selection, where the probability of "illegal" actions shrinks very significantly throughout the training process, but never reaches zero. While an attempt was made to manually remove these instances during the testing process, it is possible there may have been an inflation of upside performance. More advanced approaches such as chance constrained DRL offers potential solution for safety critical engineering systems design [221]. There may be logistical constraints which make implementation of DRL solutions in real life more complex or more costly than it appears. Furthermore, the dynamic and adaptable nature of DRL decision making is not as readily understood in terms of an actual project's implementation. The policies produced may not be fully generalizable, and the question of how best to incorporate computational insights into the design process and system operations remains an open question for future research.

To evaluate results limitation, an important issue is whether the DRL agent overfits the environment's characterization. If this occurs, it may exhibit a significant drop in performance as soon as stochastic parameters (e.g., volatility σ, drift μ) are changed, as has been recorded in other studies [222]. Out of sample testing is done to assess the potential magnitude of this effect, where the agent trained on the parameters and assumptions presented thus far is tested using scenarios sampled from different distributions i.e., using σ_{low} = 8% and σ_{high} = 25% in (3.11)-(3.12). This is equivalent to creating two additional separate testing environments, with their respective model parameters.

Table 7 captures the percentage deviation from the results obtained using $\sigma_{central}$ = 16.7%, shown in Table 6. For the case with σ_{low} = 8%, ENPV is improved for all designs, and the greatest increments come from the DRL models. VoF worsens for the FDR design – a typical observation when a flexible system faces less volatility, since there is less uncertainty to deal with – but it improves for both DRL models. This observation is striking, but also seems counter to the literature on real options analysis. The case with σ_{high} = 25% suggests there is an interesting interaction between the ability to adapt

dynamically and training a policy using DRL. VoF improves for the FDR model due to increased volatility – the expected flip side of the observation above – but it worsens for both DRL models. This may be because with higher volatility, the scenarios are too different from those encountered during training, and the optimal adaptation strategy is less efficient.

The significant improvements in VoF for both DRL models in the low volatility case, however, suggest that the policy learned during training can generalize well for *some* level of unseen conditions (i.e., lower volatility scenarios), and that the agent is able to extract additional VoF from such conditions. The results suggest, however, that the generalizability of the policy could be related to the degree of uncertainty experienced in the testing conditions, compared to that found during training. Nevertheless, this result is interesting as it illustrates the tradeoff between potential value gained from having the ability to adapt, and the limitations of training based on a particular set of scenarios. Further studies may help understand when a particular model no longer generalizes well to unseen conditions or level of uncertainty.

Table 7:Out of sample testing of performance of design alternatives considered in this study.

Case	**Metric**	**(1) Centralized Inflexible**	**(2) Decentralized FDR**	**(3) DRL-LF**	**(4) DRL-HF**
$\boldsymbol{\sigma_{low}}$ **= 8%**	ENPV	+5.6%	+3.8%	+8.6%	+6.4%
	VoF	0%	-2.3%	+15.0%	+7.6%
	VaR	+37.9%	+41.4%	+54.6%	+14.3%
	VaG	-11.8%	-5.2%	-3.1%	-5.9%
$\boldsymbol{\sigma_{high}}$ **= 25%**	ENPV	-11.6%	-6.3%	-11.4%	-8.8%
	VoF	0%	+11.7%	-10.97%	-4.73%
	VaR	-107.5%	- 26.6%	-59.6%	-34.6%
	VaG	+8.5%	+4.1%	+2.3%	+3.4 %

It should also be noted that the limitations and sensitivity of results discussed in the above section are, to a certain degree, inherent to the fact that historical data was used to build the stochastic model in order to allow benchmarking with previous results. Future work should explore the performance of this approach when completely unforeseen events and disruptions are considered, for instance using jump diffusion models, or disruptions that are manually generated. While recent studies suggest that flexibility may play an important role in dealing with unforeseen events [11,209,223], more research is needed to evaluate how the proposed approach fares under such conditions.

5.5 Computational Insights for Decision Making

5.5.1 Capacity Evolution Example Scenario

Figure 5.7 compares capacity evolution for an example distributed demand scenario for the optimal Decentralized FDR, DRL-LF and DRL-HF designs. The scenario is characterized by unusually high waste generation in sector 3 (S3), and in sector 4 (S4). The FDR model yields expansions in sectors S6 in year 2 and in S1 in year 6. The DRL-LF design instead can recognize the excess demand in S3 and expands there in year 3, while expanding in S6 later in year 7. This is a better expansion strategy than for the FDR model, as the difference between sectors S3 and S6 demand outweighs the extra transportation cost in early years. The DRL-HF model matches demand and capacity even more closely, with incremental expansion decisions in sectors S1, S3, S4 and S6 at various stages of the project when the action is estimated to return the highest expected value. Given more dynamic policies and this specific food waste scenario across S1-S6, the project NPV increases from S$31.5 Million for the Decentralized FDR design to S$32.6 Million and S$34.8 Million for the DRL-LF/HF designs, respectively.

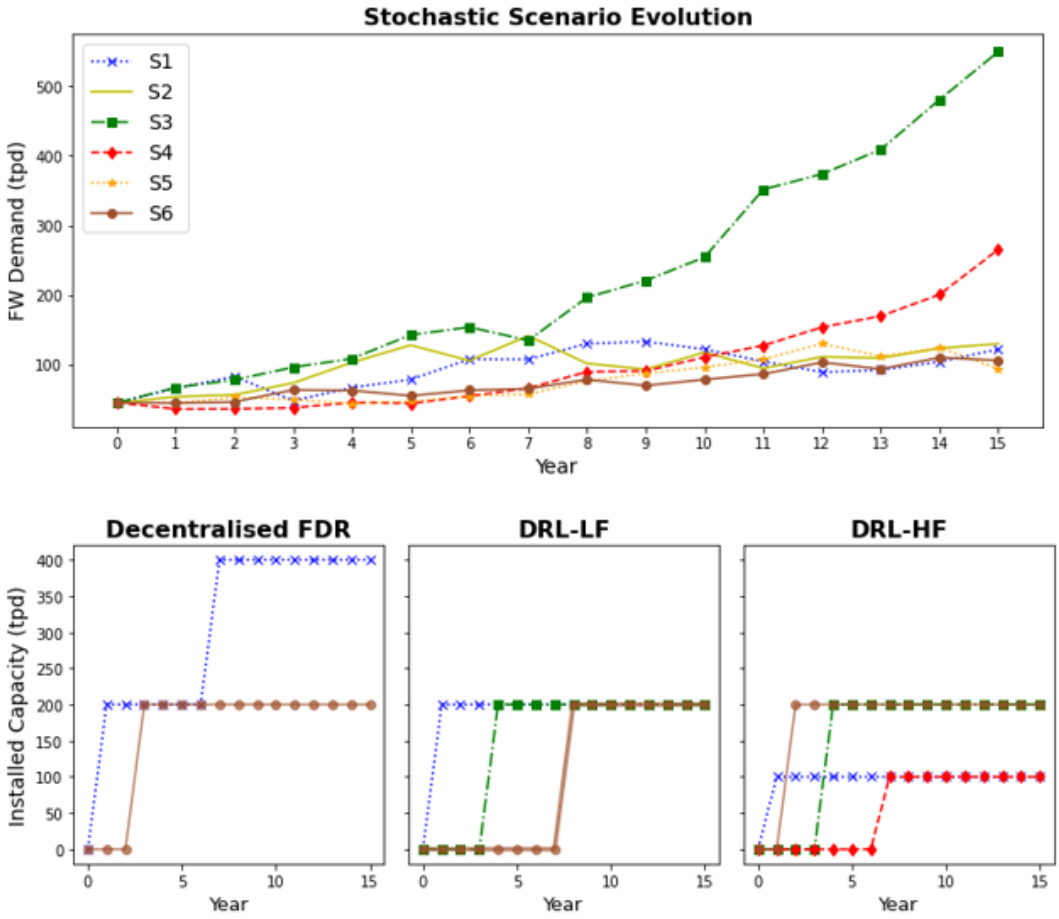


Figure 5.7: Example distributed food waste scenario leading to demand scenarios in sectors S1-S6 (top), corresponding design capacity expansion for FDR (bottom left), DRL-LF (bottom center) and DRL-HF designs (bottom right). Legend in top plot applies to all items.

5.5.2 DRL Agent Action Probability Distribution Evolution

Figure 5.8 gives a visual representation of the decision-making process for the DRL-LF agent over time, as a function of sampling probability for different actions, using the same distributed demand scenario from Figure 5.7. Following Algorithm 2, the sampling probabilities for a specific action reduces significantly once that action has been selected, or if another capacity expansion decision is executed, as it is unlikely that two expansion decisions in consecutive time periods are made. Gradual changes are normally attributed to additional uncertainty realizations made in each year as the project is underway. Using year 2 observation as input, the FDR model follows the top nodes from Figure 5.2 as Procedure 1 and 2 are *true*, leading to an expansion decision in sector S6 (see Figure 5.7). When the same state observation is given as an input to the DRL agent, following the logical progression from Figure 5.3, the most likely action returned is *no expansion* with a 65% probability. Instead, the expansion probability rises significantly for sector S3 (41%) in year 3 based on observed increasing demand, leading to that action being selected. This suggests that the agent can consider the problem more holistically, as while the per sector transport costs are highest for sector S6 that year, it may estimate that total costs in the long run will be lower if expanding in sector S3, while continuing transportation from S6. In most years, transport costs averaged 3-5% of total system cost, thus basing the expansion decision solely on that may also miss out on some potential value adding flexibilities.

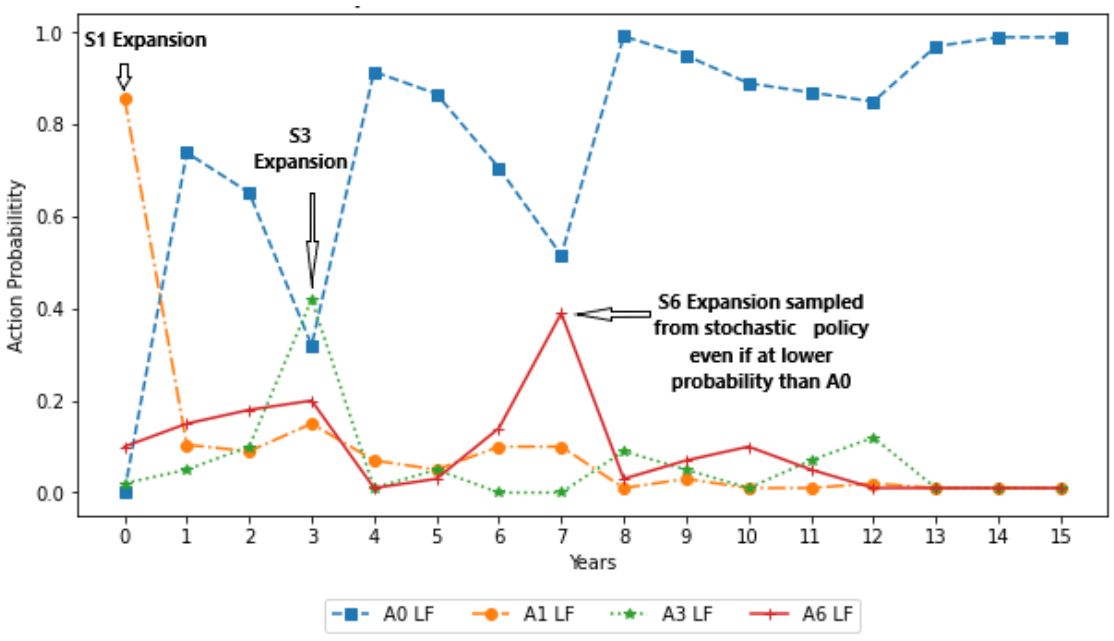


Figure 5.8: Action sampling probabilities for the example distributed demand scenario in this section, for DRL LF system. Actions maintaining a probability of being selected below 5 % through the project duration are not included.

In real-world system operations, a decision maker would be able to refer to the optimal policy and access the associated action space probabilistic distribution based on state observations (see Figure 5.8). Looking at year 7, about halfway through the project, can better highlight how this distribution of actions could be used to fit the risk criteria of individual engineering design projects and influence

the choice to deviate, or continue with the deterministic policy. For the FDR model, the system is not able to consider expansion, as the max installed capacity of 600 tpd has already been met. The DRL-LF system also suggests no expansion (i.e., action A0) as most likely (51%), however also presents expansion in sector 6 as feasible (39% probability) – and it ultimately chooses it for this episode due to the stochastic sampling of the policy. Risk-tolerance information, however, could help guide this sampling process (or deviation from deterministic policy) in a more systematic and value enhancing manner, as long as the difference in probability between two or more actions is relatively small. In this case, a more risk averse decision maker may decide to follow the deterministic policy (i.e., action A0), minimizing investment risk while also integrating one additional year of uncertainty realization before possibly making an expansion decision. A more risk prone decision maker, however, may recognize that proceeding with sector 6 expansion in year 7 could allow an additional year/capacity for increasing waste processing revenues compared to the deterministic policy, which does carry significant investment risk in case demand does not materialize. In this particular demand scenario, the deviation from the deterministic policy to proceed with early sector 6 expansion would lead to an increase in NPV for the project, although waiting an additional year is likely to provide better protection against possible downsides, hence the risk tolerance tradeoffs. On the other hand, in year 3, it may be more likely for a risk-averse decision maker to choose to diverge from the deterministic policy (A3) and select no expansion (A0) in light of their very similar sampling probability (note that this was not the case in this episode), for the same reasons discussed above. Nonetheless, there is one obvious DRL action recommendation for the majority of system operation, thus if some risk tolerance information is available for decision making, possible deviation from the deterministic policy should only be considered for years 3 or 7 in this case, as shown in Figure 5.8.

5.6 Discussion

5.6.1 Relative Performance

The NPV improvement in any given scenario occurs because of better transportation cost management as compared to a Centralized Inflexible design. It is also due to exercising the expansion options at a more optimal time compared to a Decentralized FDR design. NPV instances greater than S$39 Million for the DRL-LF design were removed during ENPV calculations in Table 1 since they violated maximum installed capacity of 600 tpd, due to the stochastic nature of the policy. As this would allow them to capture more upside potential compared to alternative solutions, the choice of removing them was made to allow for more appropriate benchmarking against other methods. Nonetheless, this represents only the very tail end of the distribution (less than 1% of simulated scenarios), and thus Figure 5.6 is still representative of their relative performance.

It is observed that the DRL-LF design yields worse downside performance and overall variability per episode (equivalent to one 15-year simulation) compared to the FDR model. These indicators create valid concerns and suggest further fine tuning of the solutions may be needed to increase acceptability

in a real-world setting, where VaR is an important metric for decision-making. The larger capacity expansion flexibility allowed by the DRL implementation may favor looking for upsides at the cost of neglecting downsides, although ultimately the effect of increased upside potential remains dominant. Including flexibility strategies more focused on addressing downside risks, such as abandonment or temporary shutdowns, could potentially aid in mitigating these effects, and yield a narrower CDF profile. This kind of trade-off is something that could be determined based on individual decision makers preference during design space exploration activities.

Nonetheless, there are certainly scenarios where the flexible solution is not guaranteed to be stochastically dominant relative to the non-flexible solution, given the associated costs of flexibility, and form of the design problem itself, highlighted by the long downside tail for DRL-LF in Figure 5.6. Visual inspection of DRL-LF/DRL-HF policy and scenarios suggests that this may be due to simulations with relatively higher initial demand and relatively much lower growth rates over time in specific sectors. This can lead the DRL-LF agent to overbuild in certain sectors in early years at a significantly higher cost than for the inflexible design, possibly based on overestimations of future demand evolution within the neural network policy, while also incurring higher transportation costs. Interestingly, the more modular capacity expansion allowed in DRL-HF is much less sensitive to these cases, as smaller capacities are normally deployed in early years in these scenarios. This helps to spread out the investment risk over time and allows the episode policy to be progressively updated according to the uncertainty realized, limiting the instances of uncorrectable overbuilding in certain sectors and associated downside tail magnitude compared to DRL-LF.

Finally, the results on this case study suggest that the proposition of flexibility as a unifying paradigm for sustainability and resilience under uncertainty, introduced in 1.2, is justified. In this engineering systems design problem, the flexible solutions (and particularly the ones generated with the proposed approach) yield lower operational emissions due to a reduction in transport requirements. While the emissions reductions are not directly estimated here, the connection seems apparent as a the most modular and dynamic solutions lead to lower fuel consumed for waste collection. The result is an improvement in the overall sustainability of the engineering system design as compared to benchmarks or alternative methods. Furthermore, while only assessed in terms economic terms, the reduction in downside exposure produced by the proposed methodology (i.e., improvement in VaR) suggests the development of a more financially resilient design under uncertainty.

5.6.2 Degree of Embedded Flexibility

In the DRL-LF implementation, the primary insight is the optimal triggering signal to exercise flexibility, as capacity expansion decisions are limited to the level obtained from the FDR approach. In terms of starting capacity, the same conclusion is reached as in the FDR model, to install the initial capacity of 200 tpd in Sector 1 (main site). This is a result of this decision being the most likely to lead to the lowest transportation cost over system lifetime based on the assumptions from Table 14 in

11.3. In terms of optimal capacity expansion strategy, similar conclusions are also reached. The FDR approach yields expansion in sectors S1 and S6 for most simulated scenarios, resulting from those sites presenting the highest collection distance and thus transportation cost for undersized capacity. Exceptionally high demand scenarios in sectors S4 and S5 do occasionally lead to expansions within their area, but this is often constrained by the form and resulting timing of the decision rules. The DRL-LF approach presents a broader distribution of possible actions. Other than no expansion (A0), the highest probability is found for expansion in sector S1, followed by sectors S6, S5, S4, S3 and S2 in that order, which seems nearly optimal given the assumptions presented in Table 14.

Evaluating the DRL-HF design yields further insights and an increased value of flexibility compared to the DRL-LF implementation. In terms of starting capacity, the optimal suggested rule is between 100-200 tpd of initial capacity installed at the main site. The findings on expansion magnitude primarily confirm the results from the decision rule approach as the highest probabilities are generally found for 200 tpd action for each sector, to take best advantage of economies of scale. The magnitude of the EoS factor is an important determinant, and it is expected that for reduced EoS the probability of selecting smaller, more modular expansions would increase. In terms of expansion decision distribution, a visual inspection of past actions distribution shows that unlike in the DRL-LF implementation, the solution can likely be further optimized. The primary indicator is a relatively high probability of selecting action A5 (expand by 100 tpd in Sector S2) compared to the DRL-LF implementation. Aside from cases of extremely high demand in S2, this is unlikely to be the optimal decision as more significant reductions in transportation costs can be achieved by expanding in other sectors. This suggests that the DRL-HF agent may be somewhat stuck in a local optimum, although more training and different initialization conditions offer potential solutions to this issue.

5.6.3 Implementation during Engineering System Operation

In a practical setting, in any given year, the system state (capacity and demand in each sector) could become an input to the optimal policy, which would return an action probability distribution as introduced in Figure 5.3. This distribution can be used as a proxy for evaluating the estimated expected value of different actions. Having access to this kind of distribution rather than a single recommended action obtained via FDRs, could allow decision makers to integrate their own expertise and insights into policy outputs for choosing system operation strategy, increasing the applicability to real world situations. The maximum relative difference in probability of selecting different actions to justify a potential deviation from the deterministic policy is something that will vary among engineering systems, based on associated uncertainty and action space size, with potential adjustments through decision makers risk profile and project specific insights. Future work could focus on determining the appropriate thresholds to warrant considering deviation from the deterministic policy, and the resulting implications for system operation. In most decision time-steps, however, it is still

expected that the deterministic policy would be followed, with decision makers progressively updating system state description year to year and implementing the resulting recommended action.

Furthermore, say an important but possibly uncertain parameter like fuel cost changes significantly by year 2, the DRL agent can observe the change, and dynamically adjust the policy to match the new environment based on a few computationally inexpensive interactions. This is often not the case in standard methods to analyze flexibility, where initial assumptions are usually maintained throughout implementation to compute stochastically optimal decision rules – although approaches such as post-optimality sensitivity analysis can help determine the thresholds for which these solutions break down. These examples and the scenario presented in 5.5.1 give some intuition on how the action distribution changes based on underlying state observations, and thus how the current state of the system could be used as an input into the DRL policy at various stages, helping to determine preferred operation strategy.

5.7 Chapter Summary and Conclusions

This chapter extends the novel proposed methodology based on DRL to complement existing approaches to analyze flexibility in engineering systems design. The proposed approach helps explore alternative solutions or configurations through thorough a systematic data-driven analysis of the solution space, uncovering alternatives that may not be considered using standard design methods. A second example implementation is shown on the design and analysis of a waste-to-energy (WTE) system in Singapore, with capacity expansion flexibility across spatial and temporal dimensions. The results confirm that embedding flexibility in engineering systems under uncertainty improves expected performance significantly as compared to stochastically optimal, but inflexible designs. Furthermore, they show that the DRL approach produces highly adaptable and dynamic design and decision rules to navigate an uncertain environments, even for highly complex engineering systems, thereby further improving the solutions from a previous study across all metrics evaluated [208].

Building upon the approach based on decision-rules and value-based DRL presented in the proof-of-concept Chapter 4, the contributions of this thesis to the field of Flexibility in Design were extended here in a few crucial ways. Firstly, the results presented showed the significant potential of the novel data-driven approach for highly complex and dimensional case studies, an important motivator for its development. This was particularly highlighted by the performance improvements for the higher flexibility (DRL-HF) design and strategy evaluated here. The dimensionality of such a system would likely lead to a computationally intractable optimization using standard methods, thus marking an expansion of the feasible design space which may be considered. Secondly, a novel way to implement the proposed approach in practice is explored, by using the underlying action probability distribution returned from a given state observation. A more simplistic statistical analysis of decision rules implemented is also presented, although this only really served to gain computational insights for

decision-making interpretability, and thus does not represent a practical triggering signal. Finally, to better assess claims of improved generalizability though NN function approximation, the performance of designs produced via different methods was assessed when changing stochastic model parameters. This showed that the proposed method can generalize well to some unseen scenarios, highlighting an important trade-off between generalizability, computational expense, and performance under the most likely uncertainty parameters.

There are several limitations that can be addressed through future research. Future efforts could investigate integrating additional flexibility (i.e., all generic real options) and uncertainty sources (i.e., fuel costs, waste purity rate, etc.) into the analysis, to help determine the problem dimensionality for which this approach is no longer suitable. Increased integration of year 0 design decisions, accounting for technical enablers, could help uncover further value for the proposed approach. Multi-objective reward functions, looking at sustainability of the whole project through ESG metrics, for example, could also be implemented with some very important operational insights. The introduction of random "shocks" or perturbations to environment dynamics could help investigate the potential for DRL to design more resilient engineering systems. More advanced statistical analysis of agent actions under different conditions could also help to clarify operational insights and reformulation of the modeling outputs as usable guidelines. In this chapter, a different approach to that taken in the proof-of-concept case study was implemented to gain computational insights for practical decision-making, although it is difficult to determine which is better suited in the lack of stakeholder interaction. Finally, it would be interesting to understand what kind of environments and design problems are better suited for a stochastic rather than deterministic sampling policy on action space (i.e., select the highest probability), and the resulting implications on system performance.

In order to further mature the novel approach to analysing flexibility in engineering systems proposed in this thesis, therefore, it is deemed appropriate to implement it on an even more complex, dimensional, and uncertain case study. This would help to further stress-test the original motivations, and determine is such an analysis can be performed for engineering systems which would otherwise be computationally prohibitive. Additionally, it may be useful to directly assess (and perhaps optimize) sustainability and resilience metrics, gaining further insights on the effectiveness of flexibility to complement these crucial objectives.

Chapter 6

6 Methodological Maturation: Mobile Micro-Grid Design and Planning

This chapter marks a significant maturation of the proposed methodology, by presenting a successful implementation on an extremely high dimensional application case study. Generally, the analysis presented here was driven by the following question.

- RQ5: What are the computational limitations of DRL in terms of case study complexity and can it handle design problems that have never been tackled with alternative methods*?*

More specifically, this chapter shows how a data-driven approach to analyse flexibility in engineering systems design allows for the first time the planning of a mobile micro-grid system enabled by recent distributed control developments. As such, it is a very important contribution to energy access for global nomadic communities, and marks a substantial shift from standard methods used in the field of energy systems planning. The significant number of flexibility enablers, as well as operational and planning uncertainties, help highlight the added value through the method proposed in this thesis as compared to standard approaches. It should be noted that the novelty of the work presented in this chapter, with its associated uncertainty space, does not allow for direct benchmarking against alternative state of the art methods (or other previously published results). The consistent performance improvements obtained for the case studies presented in Chapters 4 and 5 justify the lack of extensive benchmarking with standard methods here. Several configurations of inflexible baseline designs are, nonetheless, optimised for relative performance evaluation. As such, no comparison with pre-defined decision rules is presented, and a novel approach based on advanced statistical decomposition of decision making under different conditions is explored. This was targeted to be more generalizable than the retroactive distribution of decision-rules from Chapter 4, as novel and intuitive triggering signals are discovered, contributing to computational creativity and concept generation for flexibility strategies. Furthermore, as sustainability and resilience metrics can be directly defined for energy system operation, this case study provides a more holistic understanding on DRL Flexibility analysis as the unifying paradigm introduced in 1.2. A prototype for a design decision support tool around the proposed approach is also introduced here. In summary, this chapter addresses the key objectives and limitations from the methodological validation and expansion sections, showing how the methods in this thesis may potentially be applied to the design of any engineering system under uncertainty.

Most of the work presented here is a result on the Global Challenges Research Fund (GCRF) project : "Energy for Development – low carbon energy and industry for economic growth in Mongolia", pursued in collaboration with the Energy Futures Lab at Imperial College London. In particular, Dr. Fei Teng and Pudong Ge were responsible for the development of the novel Plug and Play (PP) control algorithm which enables this kind of analysis from a technical energy control perspective. Dr. Anna Korre, on the other hand, contributed invaluably to discussions on capturing system definition and boundaries as well as considering the impact and design of different policy mechanisms. Finally, Alex Gibson led the development of the decision support tool prototype, under my close supervision.

The results in this Chapter are primarily based on the published peer-reviewed journal paper[134] and a series of collaborative workshops held with Mongolian energy sectors stakeholders as part of the GCRF project[199]. The full open source code to reproduce all results presented in this chapter can be accessed here[217].

6.1.1 Introduction

6.1.1.1 Climate Change and Migration

The climate crisis experienced in recent years is inducing significant changes in temperature, precipitation, and extreme weather event patterns 198. This is strongly impacting the livelihood of nomadic communities, which is highly dependent on local environmental conditions. Nomadic communities may be generally viewed as "ecologically adjusted at a particular technological level to the utilization of marginal resources in areas unsuitable for sedentary agriculture" 199. As such, most of them are dependent on following an approximately fixed seasonal pattern of movements to maintain their livelihoods in arid or semi-arid conditions. Movements are traditionally performed using livestock, transporting some form of mobile dwelling in the pursuit of pastures which can better sustain agricultural activities. It is estimated that there are currently a least 30-40 million rural nomads, primarily found across Sub-Saharan Africa, South/Central-Asia and Central America 200.

On the other hand, areas where sedentary agriculture can be sustainably practiced are rapidly decreasing, leading to climate-change induced mobility worldwide 201. Recent models suggest that around 143 million more people will be displaced by 2050, creating a new wave of "environmental refugees" 202,203. The majority is expected to be within regions already practicing nomadism, with a greater impact felt in the poorer and more vulnerable communities. Climate driven relocations, nonetheless, are expected to happen in several developed countries as well 204,205.

Climate driven migration is also recognized as one of the major causes of urbanisation worldwide 206. Urban migrants are normally composed primarily of unskilled workers associated with the informal sector, where the lack of proper steady housing has limited their integration into the formal sector. They typically create informal settlements on the outskirts of major cities, housed in small tent-like structures. As such, they are normally unable to access services such as electricity or district

heating. The resulting energy poverty impacts the livelihood of urban migrants significantly, having to spend long periods of time and a high proportion of their disposable income in the search for more rudimentary power sources, such as dung or biomass 207. Furthermore, it was estimated that globally there are at least 208 million urban migrants completely lacking access to energy, primarily due to infrastructure limitations or lack of access to the formal government sector 60,206.

While the UN Sustainable Development Goals (SDGs) do not explicitly link climate change and migration, it is becoming increasingly important to develop policies and technologies which help mitigate the impacts of climate induced mobility 208. Energy access is very complex in this context, as the current infrastructure is not compatible with such mass migratory movements, particularly in lower income countries 209. In order to address SDG 7 (Affordable and Clean Energy for All) it is, therefore, imperative to develop energy system solutions which are compatible with a migratory lifestyle, as opposed to the "static" approach used in energy systems planning thus far 210.

6.1.1.2 Need for Plug and Play Based Energy Systems

Given the current high costs for national grid extensions, mini-grid and micro-grid solutions are expected to be the main driver of electrification for almost half of all households currently lacking energy access 211. The rapid cost decline for renewable energy systems (RES) over the past few years has already increased their rate of deployment, providing electricity to 140 million people as of 2020 60. On the other hand, there are ongoing trends towards more distributed generators (DG) as RES are increasingly installed closer to their respective loads, reducing transmission power losses as opposed to remote centralized generation. The uncertainties and intermittence of RES generation, however, lead to several technical challenges for their integration into the main grid. At higher levels of RES penetration in particular, maintaining operational parameters such as voltage and frequency within their allowable limits can become increasingly problematic 212.

Nomadism represents a more extreme example of the above issues, as both loads, and generation are not only distributed but also changing over different temporal and spatial scales. Simultaneously, remote working policies implemented as a result of the COVID-19 pandemic have led to a global increase in "digital nomads" 213. As a lot of these individuals move around in campers or recreational vehicles for various reasons, they also contribute to increasingly decentralized and spatially unpredictable electricity generation and demand. Based on the above and the trends mentioned in 6.1.1, there is a growing need for tools to analyse the techno-economic potential of flexible and mobile energy systems - which are able to switch operation between a centralized and decentralized manner to accommodate both rural and urban migrants throughout their movements.

Recent advancements in distributed control algorithms have enabled the consideration of mobile and decentralized energy system solutions in the context with a migratory lifestyle 214,215. In most regions currently practicing nomadism, in fact, family clusters tend to be relatively close to each

other, following similar seasonal migration patterns 200. This creates the possibility for temporary interconnection among them to form a mobile herder household energy supply system (i.e. micro-grid), or allowing flexible plug-in operations with the main grid. This sort of plug and play (PP) operation can be highly advantageous as it allows the smoothing out of load/generation imbalances among various households, with potentially different electricity generation and storage capacities. Plug-in operations with the main grid, on the other hand, could be beneficial by providing an additional energy supply to nomadic communities when available, while helping improve main grid stability through a more distributed storage capacity 216. Additionally, market engagement opportunities may be generated, helping reduce the total electrification costs for nomadic communities – particularly as opposed to a fully centralised approach.

Moreover, PP holds potential to accelerate the decarbonization of heating supply systems for nomadic communities living in colder regions. Their traditional heating systems rely on inefficient combustion of fossil fuels and could be replaced by electric-heating (EH) technologies if a reliable and sufficient level of sustainable electrification can be achieved. This may be particularly attractive if implemented at a community rather than single household level, as a higher degree of modularity can be integrated in system capacity evolution strategies 206. At the time of writing, however, all studies on PP have focused on fine timestep resolution control operations to guarantee network reliability. Furthermore, none have analysed it through a multi-energy vector perspective to accelerate the transition to sustainable heating. Given the potential long-term implications of PP as climate induced migration increases, a novel modelling and planning framework which integrates mobility considerations is needed to quantify its long-term impact, and develop policy recommendations – for both rural and urban migrants.

PP operation can be allowed without control law changes in the power-electronic-level devices, guaranteeing system stability and synchronisation (in terms of frequency and voltage), as detailed in 6.2.2. As such, it is compatible with operational control strategies including predictive Machine Learning (ML) methods, advanced control algorithms, or heuristic-based approaches 212,217-221. Nonetheless, recent studies suggest that for smaller scale systems with no dispatchable generator, a simple Load-Following Strategy (LFS) is likely to result in similar conclusions on initial design capacities and total costs as more complex methods 221-223. Its logic prioritises renewable generation, followed by battery energy storage and grid interactions, if available, to meet any remaining load/demand – adopting simple heuristics to operate the system. Accordingly, it is well suited to the technological and communication capacity limitations associated with a nomadic mobile energy supply system. Based on the more limited control opportunities available, a flexible capacity planning strategy under uncertainty is, therefore, more likely to impact the distribution of total system costs (as well as other performance metrics) obtained through long-term PP operation.

6.1.1.3 Energy System Planning and Role of Flexibility

The literature on general micro-grid planning over a long horizon for rural electrification is very extensive, although yet to be combined with PP enabled mobility considerations 224-227. The majority of "static" design approaches are based on either Mixed-Integer-Linear Programming, Multi-Stage Stochastic Programming (MSSP), or evolutionary methods such as the Genetic Algorithm (GA) 226. The aim is typically minimisation of Net Present Cost (NPC), although some works more closely looked into sustainability 228,229 and resilience 215,230,231 considerations too. In order to reduce the computational expense and obtain an analytical solution, however, significant simplifications tend to be made as opposed to the real-world problem. The study in 232, for instance, only used a single "representative" year to determine system performance over the whole project horizon. As shown by 233,234, however, the inability to account for load growth or corresponding energy system evolution limits the applicability of single year approaches in an energy access context. Consequently, uncertainties related to future load growth were previously found to be significantly more impactful on final system costs than those related to RES generation 226. Other works tend to focus on reducing uncertainty recognition to only one or two sources, with their realizations limited to a predefined set of scenarios 235. The authors in 236, for example, develop stochastic models for load growth but not for renewable electricity generation throughout the project horizon. Other common simplifications include considering only a specific technological component (i.e., battery in 237 for example) or using just a few representative days for each year 226,227. Furthermore, only one level of expansion is normally considered for each technology, with limited generalizability and adaptability outside of projected scenarios 222,223,238.

The planning and design of a nomadic energy supply system, therefore, must balance delicate trade-offs between mobility, security, efficiency, sustainability, and economic feasibility in light of the extensive associated operational uncertainties. . Flexibility is a potential value enhancing paradigm to help address some of the above-mentioned limitations of common energy system planning approaches 48,54. Planning a PP based mobile energy system through a Flexibility in Design approach can, as a result, likely help mitigate investment risk and fiscal burdens incurred by the targeted low-income nomadic communities under uncertainty. Standard methods for Flexibility in Design, however, present many of the same computational limitations as the planning approaches mentioned in the former section, motivating the work presented in this thesis.

The proposed DRL based approach to analysing flexibility allows for the tractable integration of several sources of uncertainties and mobility considerations associated with a mobile energy supply system. In the energy systems field, DRL has been extensively investigated for management and control 239, but it has yet to be directly implemented on micro-grid design and capacity planning for rural electrification. The closest work would be that by 240, however DRL was only used by the authors for the dispatch problem, and components were sized through evolutionary methods.

Additionally, energy system capacity planning or decision-making flexibility under uncertainty were not considered. The design and uncertainty space relevant to an energy system planning problem, it follows, can be significantly expanded as opposed to analytical methods such as MSSP 10. Consequently, a data-driven approach may also help identify novel, unintuitive and more dynamic adaptation strategies outside of the standard generic real options framework. This is particularly valuable in the case of a unique, complex, and highly uncertain context such as the one tackled in this work, where there is limited domain expertise. As a result, energy system designs and flexibility strategies developed through a data-driven approach are likely more compatible with (and representative of) the range of real-world conditions potentially faced by mobile energy systems in the future.

6.1.1.4 Chapter Contributions

In light of the pressing issues discussed above, this chapter presents the design and planning of a mobile multi-energy supply system which is compatible with a nomadic lifestyle, seeking to integrate both electrification and heating supply decarbonisation objectives. To address deficiencies of common "static" energy planning approaches in this context, a new modelling framework is proposed where the spatial dispersion (and network configuration) among nomadic dwellings varies based on seasonal requirements, as well as main grid availability. In order to tackle the resulting highly complex and dimensional stochastic optimisation problem, a novel DRL based data-driven approach is proposed to avoid the simplifications and compromises associated with typical planning methods. The primary contributions are summarized below.

1. Development of a novel modelling framework for the design and planning of mobile multi-energy supply systems, focused on achieving SDG7 for global nomadic communities, but applicable to general trends in increasing decentralization of energy systems. This represents the first study integrating considerations of seasonal centralized to decentralized operation over a long planning horizon, enabled by PP operation. To the authors knowledge, the novel approach proposed is the first allowing for changing network topologies and/or spatial dispersion over time, while capturing the fiscal and mobility constraints faced by a typical targeted nomadic community. The framework is developed to be generic and easily scalable to the design and capacity planning optimisation of different communities wishing to operate a mobile energy supply system.

2. Implementation of a new DRL based approach optimising long-term flexibility strategies for nomadic communities, marking the first time DRL is used in an energy access planning context. Several sources of uncertainty are, as a result, able to be tractably integrated, including on energy generation and demand, spatial dispersion, and main grid availability. This approach enables for the first time the consideration of monthly adaptability strategies, capturing the interactions between mobility and seasonality on optimal system design. Resulting policies can be highly dynamic, and are more generalizable than those obtained using only a few select scenarios. Results are benchmarked

against baseline rigid designs with better economies of scale to estimate the values of decision-making flexibility in the context of a migratory lifestyle across economic, sustainability and resilience metrics.

3. Multi-objective assessment of value of PP operation and "breakeven" spatial dispersion point to determine investment guidelines and policy insights to accelerate nomadic energy access. This is performed using a variation of real-options theory and represents the first study focusing on long term planning of mobile PP enabled energy systems enabled. As such, it yields important quantitative results on its practical feasibility and global impact potential.

The novel modelling framework and methodology is exemplified through a case study based on Mongolian herder communities, developed in collaboration with local energy sector stakeholders. The general approach, however, may be easily extended to other nomadic communities worldwide, or to the overall needs to better understand the feasibility of mobile energy systems which switch periodically between centralised and decentralised operation to increase network reliability.

6.2 Case Study Overview

6.2.1 Mobile Micro-Grid Mathematical Model Formulation

6.2.1.1 Problem Statement

This section describes the formulation of the model for the planning of mobile-multi energy networks for nomadic communities over a long-term horizon. The integrated model accounts for changing and uncertain movement patterns over time, micro-grid distribution losses, variation in renewable energy output, and different heating generation sources. Furthermore, it considers the possibility of connection/ market engagement with the main distribution grid when available, allowing switching between centralized and decentralized operation based on seasonality and user requirements. Hourly resolution is used for power flow calculations to tractably approximate generation, distribution, and losses at the system level. On the other hand, monthly resolution is implemented for cost calculations within the flexible capacity planning model. This allows for adaptation strategies based on changing climatic conditions and intra-year seasonality, while considering the impact of long-term realisations of uncertainty on mobile micro-grid performance.

Two different mobility scenarios (MS) are thus developed to reflect the diversity in movement pattern. MS1 depicts operation of the micro-grid in the rural or decentralized setting, typically in the summer months for nomadic communities as they migrate to maintain their livestock. MS2, on the other hand, represents operation in a centralized setting, or when connection to the main grid is available, typically in winter months for nomads - when pasture productivity is lowest.

Two primary alternatives are compared based on the model presented in the following sections for a community operated mobile micro-grid. Firstly, an inflexible baseline is considered where all capacity is installed at project start and no adaptation strategies in response to uncertainty are possible.

Secondly, a flexible design exercising modular deployment of different technologies over time based on realised operating conditions is developed. The baseline rigid design presents lower unit costs due to economies of scale and reduced transport charges, while the flexible design allows for the deployment of resources only if and when needed. The objective for both designs is minimization of Expected Net Present Costs ($ENPC_{es}$) for the energy system (es) over 30 years (T).

Nonetheless, this work is also motivated by the UN SDGs and associated sustainability (i.e., Scope 1 and 2 emissions) and reliability energy access objectives. Emissions are thus found from the direct combustion of coal for heating and environmental footprint of the electricity purchased from the main distribution grid, respectively [224]. Additionally, 30-year horizon total unmet load (TUL) is used as a proxy for estimating system resilience and energy security, considering at a higher level the impact of demand and generation uncertainty. It should be noted that these metrics are not explicitly optimised, but are used for the qualitative evaluation of design alternatives in 1.1.

6.2.1.2 PP Control and Operation

A core enabler of this kind of analysis and system design methodology is the novel potential for PP operation, one of the key functionalities benefitting from the high-penetrated power electronic devices, e.g., inverter. PP operation enables the mobility of energy sources by implementing intelligent algorithms to guarantee operational stability, which is essential for compatibility with a nomadic or migratory lifestyle [225,226]. This creates an opportunity for economic, sustainability and energy security benefits, even in the rigid baseline design, as further discussed in 0. The concept of PP operations only requires real-time power-electronic level control design that has been widely investigated in [227,228]. The control objective within the micro-grid system is to guarantee that frequency and voltage converge to their reference values (i.e., frequency of 50 Hz and voltage of 220√2 V) to maintain network stability. This is achieved through active power sharing across DGs connected to the microgrid energy supply system. The images on the left of Figure 6.1 demonstrate the changing microgrid configurations and respective estimated time resolutions for a typical operational scenario encountered. Starting from normal conditions (1), one of the DGs is disconnected in (2) as a family decides to migrate, and it eventually reconnects to the microgrid in (3). The right sub-plots (a)-(d) in Figure 6.1 demonstrate that PP operation is able to maintain appropriate values for frequency $f(Hz)$, active power ratio mP, active power $P(kW)$, and voltage $V_{mag}(V)$, respectively, across scenarios (1)-(3). Over the simulated 5 second horizon, it should be noted, some minor transients are still recorded at a fine resolution due to the changing system loads. Nonetheless, the demonstration results suggest PP operation should be an effective and reliable solution to allow temporary interconnection among nomadic community members. Finally, the added value of flexibility from PP VoF_{PP} is defined according to (3.15) in this context, or the max that should be paid for the infrastructure investments required to enable this novel technological development.

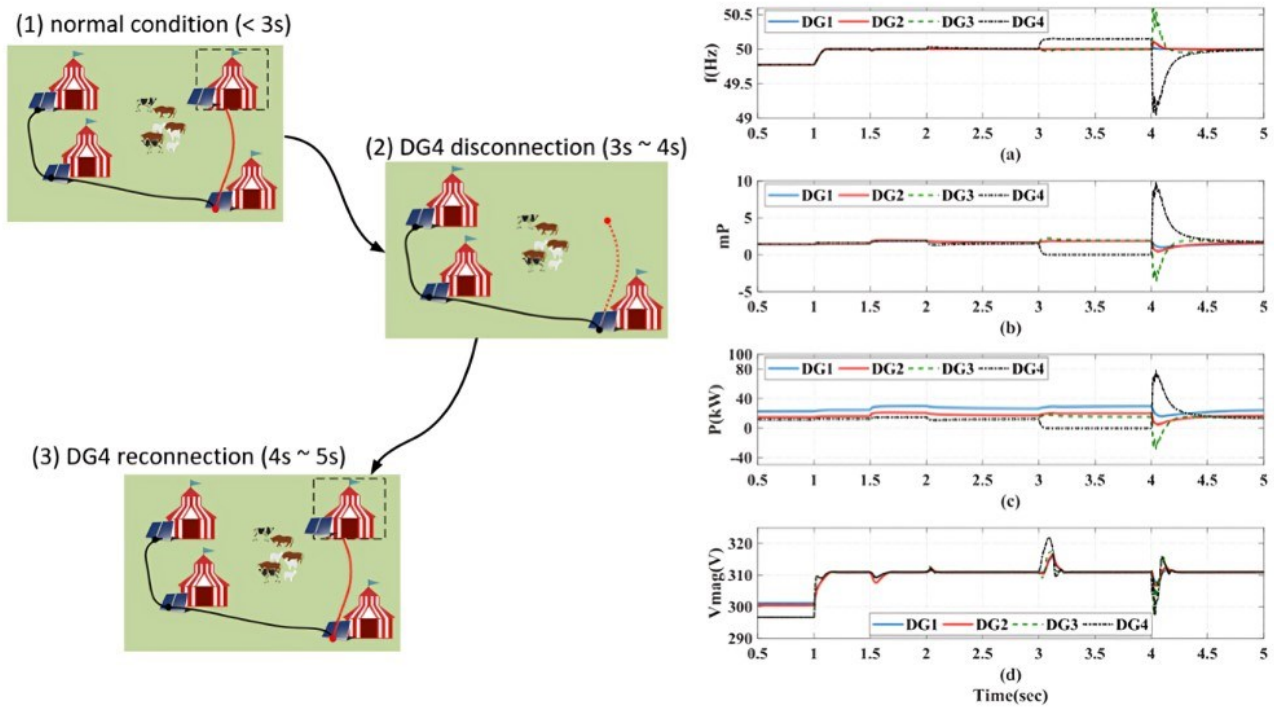


Figure 6.1: PP operation demonstration including typical (1) ~ (3) operational scenarios (left) and performance evaluation of frequency, active power ratio, active power, and voltage, in (a)-(d), respectively, over the same time horizon (right).

6.2.1.3 Nanogrid and Network Configuration

The architecture of the energy system proposed in this paper is a distributed form of mobile micro-grid. A single herder family *nanogrid (i.e. sub-microgrid) unit* (N_n) which moves and settles very close to each other is used as a building block for the simulation, allowing for improved scalability and compatibility with PP operation [229]. The term nanogrid is justified as each herder family unit is normally expected to be able to produce, store and distribute electricity in fully self-sustaining or islanded manner- as sometimes necessary during migration. Through interconnection with other community members, however, the aim is to reduce the total system cost by more efficiently managing supply and demand. Figure 6.2 graphically captures the network configuration, both at the level of an individual herder family nanogrid unit (N_n-left) and the PP mobile micro-grid (right).

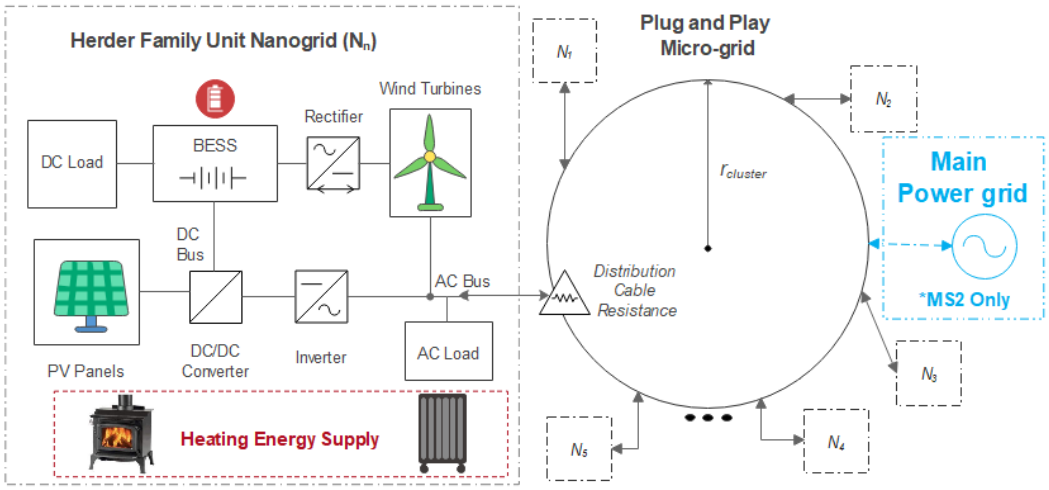


Figure 6.2: Multi energy vector block diagram of herder nanogrid (left) and PP microgrid (right).

The typical nanogrid unit consists of energy generating units – in this case solar PV and wind – as well as a storage system (BESS). While both DC and AC loads may be present in the system, interactions among nodes connected through the PP configuration is designed to be 3-phase AC, thus converters and inverters are included where appropriate. Heating demand may be present during the winter months which must be fully met using either a traditional coal stove (left) or electric heater (right). Moreover, it is assumed for it to be possible for a group of nomads, perhaps from the same extended family or community, to operate as a "cluster". The cluster radius variable $r_{cluster}$ is thus defined as the smallest possible circle which could encompass them all – as shown in Figure 6.2. It is estimated that this is not likely to vary significantly *within* a single season, although there may be important *year-to-year* differences based on the impacts of climate change. Stochastic realizations for a single month m are thus obtained proportionally to average seasonal values at project start, directly impacting the distance ${L_{cb}}^{t}$ required to connect the mobile energy system, as shown in (6.1).

$$ {L_{cb}}^{t} = 2\pi {r_{cluster}}^{t} \tag{6.1} $$

The different nodes are assumed to be configured in a fixed, ring-like interconnection, evenly spaced across the outer circumference obtained from $r_{cluster}$. This configuration was previously reported to optimise distribution efficiency under a similar decentralised setting as depicted above [230-234]. The proposed approach can, therefore, account for changing migration distances over time, while also integrating social considerations on spacing among family units. Conditional connection to the main grid during MS2 is approximated as an additional node in the network (blue in Figure 6.2).

6.2.1.4 Renewable Energy Generation

The renewables.ninja platform [235] is used for the simulation of wind ($P_{wind}^{n,t}$) and solar PV ($P_{pv}^{n,t}$) intermittent generation for each nanogrid n at an hourly resolution t, based on locations representative of each MS. The closest matches in terms of technical performance parameters were used since the

small-scale systems considered are not available for direct simulation within the platform. More specifically, Solar PV generation is found through the Global Solar Energy Estimator, with default tilt and azimuth ($\beta = 35°$, $\phi = 180°$), an assumed additional system loss fraction of 10%, and no tracking mechanism implemented [236]. Small scale wind generation is computed through the Virtual Wind Farm model and historical wind speeds obtained from NASA's MERRA-2 dataset, implementing bias-corrected reanalysis to increase the accuracy of prediction, at the lowest recorded height [237]. Moreover, to investigate the impact of RES intermittency, random Gaussian noise is added to the baseline deterministic profiles. Respective stochastic parameters for wind and PV during MS1/2 are found from their historical hourly generation potential [238].

6.2.1.5 Power Flow Calculations

6.2.1.5.1 Nanogrid Operation

The LFS is implemented for operational control simulation of the design alternatives evaluated. For any hourly timestep t, the net electricity demand load ${P_{NEL}}^{n,t}$ at node n is calculated from the hourly electricity demand $ED^{n,t}$ and on-site renewables generation $P_{res}^{n,t}$ (accounting for inverter and converter efficiencies). EH loads are prioritised over BESS charging, limited by installed capacity $\theta_{eh}^{n,t}$, hourly maximal output ${P_{eh,max}}^{n,t}$, and power available from total generation ${P_{eh,res}}^{n,t}$. The remaining power $P_{res,PL}^{n,t}$ and heating demand $P_{NHL}^{n,t}$, if any, dictate BESS operation at each node, which is prioritised over PP distribution to minimise distribution losses. The energy flows into the BESS capacity θ_{ba}^{t} are then determined from their technical parameters, namely charging C_{in} and discharging C_{out} rates as well as maximum nominal input $P_{ba,in,max}^{n,t}$ and output $P_{ba,out\ max}^{n,t}$ power. They are also bounded by the minimum $E_{ba,min}^{t}$ and maximum $E_{ba,max}^{t}$ storage capacities for the selected BESS. The battery storage level E_{ba}^{t} is thus computed as the sum of the stored energy from the previous timestep E_{ba}^{t-1}, storage energy flow P_{ba}^{t} and any potential additional installed storage capacity during that period $E_{ba,new}^{t}$. Note that newly purchased batteries are assumed to arrive *fully charged,* and conversion losses are accounted for via the battery roundtrip efficiency parameter η_{ba}. Remaining excess energy $P_{eg}^{n,t}$ after nanogrid operation is then used to engage in PP operation.

6.2.1.5.2 PP Micro-grid Distribution

Distribution requirements are calculated at the entire system level from each node, which are classified as either $N_{deficit}$ or $N_{surplus}$, based on the obtained energy balance. The system wide unmet load $P_{ul,es}^{t}$ and available extra generation $P_{eg,es}^{t}$ are thus used to roughly estimate micro-grid distribution requirements $P_{dist,es}^{t}$ at an hourly resolution. The micro-grid power distribution potential is limited by the capacity of the cables connecting the different units $\theta_{PP\ mg}^{t}$, and is capped at no more than 30% of the system wide electricity load $P_{dist,es,max}^{t}$, aiming to balance network reliability with increasing infrastructure cost. Accordingly, (6.2) yields the lowest value among the technical limitations.

$$P^t_{dist,es} = \min(P^t_{ul.es}, P^t_{eg,es}, \theta^t_{PPmg}, P^t_{dist,es,max}) \quad (6.2)$$

Furthermore, to tractably approximate performance, it is assumed that the distribution of power is spread *homogenously* across the micro-grid network, as both loads, and generation should be well distributed spatially in light of migration compatibility objectives. More specifically, representative power ($P^{t,n}_{dist} = \frac{P^t_{dist,es}}{N}$) and length (${L_{dist}}^{t,n} = \frac{L^t_{cb}}{N}$) requirements for PP distribution per each node n are developed for preliminary loss calculations. Dwellings are assumed to be connected via 3-phase transmission rated at 220/380 V through low-cost copper cables [225]. Interconnection resistance per node $R^{t,n}_{dist}$ is thus found from the manufacturer reported reference value and tuned to the actual operating conditions through the material temperature coefficient α_{cb}. Cable resistance is used to estimate energy system losses, $P^t_{dl,es}$ according to (6.3), neglecting reactive power and only focusing on thermal losses associated with the induced current $I^{t,n}_{dist}$ in each line. Finally, distribution loss proportion $DL\%$ is found from (6.4), seeking to capture overall network efficiency. The balance of power flows in the network after PP operation is ensured by computing the final excess electricity generation $P^t_{res,exc,pp}$, unmet electricity load $P^t_{uel,es,ppp}$ and/or unmet heating load $P^t_{uhl,es,ppp}$. These values then dictate utility grid interactions, as detailed in the following section.

$$P^t_{dl,es} = \sum_{n \in N} 3R^{t,n}_{dist} {I^{t,n}_{dist}}^2 \quad (6.3)$$

$$DL\ (\%) = 100 \frac{P^t_{dl,es}}{P^t_{dist,es}} \quad (6.4)$$

6.2.1.5.3 Utility Grid Interaction and Coal Requirements

The potential to buy electricity from the distribution grid during MS2 is considered in this section. Interconnection is simulated at a system level through simple heuristics (i.e. market engagement is *not explicitly* considered) to focus on flexibility strategies. Interactions with the main distribution grid, limited by line capacity θ^t_{Grid}, unmet heating load $P^t_{uhl,es,ppp}$, and availability ρ_{grid}, are calculated based on system energy balance, as shown by (6.5) below. Note that ρ_{grid} is a binary variable representing grid availability for that timestep, sampled from a *beta* distribution. Case (i) then represents timesteps when blackouts are happening so no interactions happen. Case (ii) shows situations when demand has not been fully met, thus electricity is purchased from the grid. For case (iii), on the other hand, the excess generation after meeting micro-grid loads $P^t_{res,exc,pp}$ is sold back to the main grid. Note that excess generation P^t_{ex} which cannot be captured by grid interactions is assumed wasted at this stage. Please refer to 11.5 for the complete power flow model equations.

$$P_{grid}^{t} = \begin{cases} 0 & \rho_{grid} = 0 \ (i) \\ \min(\theta_{Grid}^{t}, {P_{eh,max}}^{t}, P_{uhl,es,ppp}^{t}) & P_{uhl,es,ppp}^{t} \geq 0, \rho_{grid} = 1 \ (ii) \\ -\min(\theta_{Grid}^{t}, P_{res,exc,pp}^{t}) & P_{uhl,es,ppp}^{t} \leq 0 \, ; \rho_{grid} = 1 \ (iii) \end{cases} \quad (6.5)$$

6.2.1.6 Emissions Generation

Based on the above, the heating demand not met via EH, and therefore mass of coal required for the *entire nomadic community*, is calculated as shown in (6.6)-(6.7). $P_{coal,es}^{t}$ captures the thermal energy requirements after PP distribution $P_{uhl,es,ppp}^{t}$ and grid interactions P_{grid}^{t}. The associated coal mass ${M_{coal}}^{t}$ is estimated from its average energy density $\mu_{local\ coal}$ and thermal efficiency of the traditional stove used for combustion $\eta_{coal\ stove}$. Energy system carbon footprint over time is estimated from the mass of raw coal burned and grid purchased electricity according to (6.8), using their respective weighted average emissions factors (EF_{grid} and EF_{coal}).

$$P_{coal,es}^{t} = \max\ (P_{uhl,es,ppp}^{t} - P_{grid}^{t}\ , 0) \quad (6.6)$$

$${M_{coal}}^{t} = \frac{P_{coal,es}^{t}}{\eta_{coal\ stove}\, \mu_{local\ coal}} \quad (6.7)$$

$${CO_2}_{eq.emissions,es}^{t} = EF_{grid} P_{grid}^{t} + EF_{coal} {M_{coal}}^{t} \quad (6.8)$$

6.2.1.7 System Costs

The model presented thus far ultimately seeks to calculate system costs at a monthly m resolution C_{es}^{m} in each uncertainty scenario. As shown in (6.9) below, the formulation accounts for investment costs C_{inv}^{m}, operational costs C_{opex}^{m}, any expenses (or revenues) incurred through interactions with the distribution grid C_{grid}^{m}, as well any potential penalty (or credit) associated with generated emissions C_{CO2}^{m}. Representing liquidity and economic limitations faced by these communities, a constraint is formulated to cap monthly household energy system expenditure at the cluster level, as shown by (6.10). The modelling of each of these cost components is discussed in more details below.

$$C_{es}^{m} = C_{inv}^{m} + C_{opex}^{m} + C_{grid}^{m} + C_{CO2}^{m} \quad (6.9)$$

$$C_{es,max}^{m} \leq C_{es,max}^{m} N_{n} \quad (6.10)$$

6.2.1.7.1 Investment Costs

Investment costs in each timestep, as shown in (6.11), are calculated from the communication ${CC_{n,i}}^{m}$, infrastructure IC_{es}^{m} , expansion $EC_{n,i}$, and replacement costs $RC_{n,i}$, net the residual value (if any) $SV_{n,i}$ recorded across all nodes and technologies evaluated. Communication infrastructure costs CC_{es}^{m} required for the safe operation and control of the PP microgrid are preliminarily estimated through a

simplified model based on system distribution requirements $P^{m}_{dist,system}$ and a cost constant factor c_{cc}, as shown in (6.12). Assuming a 2G/3G/4G base station is already present, the communication infrastructure needed is thus simply an add-on support enabling peer-to-peer communication among nanogrid units. They are considered as a one-off investment incurred only in timesteps where $P^{m}_{dist,es}$ is higher than monthly distributed electricity in all previous periods $P^{m,max,past}_{dist,es}$, representing the need for infrastructure upgrade - neglecting the effect of EoS. Similarly, IC^{m}_{es} – which represent the need for additional cabling to connect the various nanogrid units- are recorded for timesteps when the current cluster radius is higher than that recorded in any of the previous simulation months ${r_{cluster}}^{max,past}$. It is thus assumed that cabling infrastructure purchased at any point during project life will remain available for connection if needed again. The cost of required extra cabling is computed according to (6.13), using a unit length cost factor c_{ic} obtained from regional manufacturers [239], invariant with time or scale.

The formulation of $EC_{n,i}$ in (6.14), on the other hand, accounts for both of economies of scale (EoS) and cost of transportation in the targeted settings. The estimations for the investment constant K_i and EoS factors α_i are, therefore, slightly reduced and increased respectively in MS2 as compared to the MS1 [190], and based on specifications from regional manufacturers per component [240-246]. Coal stove capacity is an exception to this, where it is assumed that no starting stage investment is required, although replacement costs are incurred every 7-years [247]. As such, stove capacity is not explicitly optimised, and it is simply assumed one will be needed per family at a fixed (and nearly negligible) cost over time. Any other given capacity decision $\theta^{m,new}_{n,i}$ is expected to yield additional power electronics requirements $\theta^{m,new}_{n,pe}$, assumed to be proportional to expansion magnitude. This is based on reported ratios of nominal power electronic to system component i capacity, $\varepsilon_{pe,i}$, averaged over several micro-grids from the literature [57,232-234,248,249]. A safety factor sf_{pe} of 1.3 is included to further ensure no undersized components, as shown in (6.15), yielding total power electronics costs ${PEC_{n,i}}^{m}$ through a constant cost factor $c_{pe,i}$.

Replacement costs $RC_{n,i}$ are incurred when a system component i reaches the end of its predicted operational life L_i, using the same parameters as (6.14). As such, $RC_{n,i}$ may be experienced for multiple technologies in the same month- in contrast to $EC_{n,i}$, which is limited to a single system component i per timestep. The asset salvage value $SV_{n,i}$ is included for cases where capacity abandonment decisions are made, or for NPC calculations at the end of each simulated episode. As shown in (6.16), it is a function of remaining estimated functional hours ${L_i}^{\theta}$ for a component with an original maximal lifetime ${L_i}^{max}$, and initial investment cost based on the salvaged capacity amount $\theta^{m}_{sv,n,i}$, corrected through a depreciation factor k^{salv}_{i} which is smaller than 1.

$$C_{inv}^{m} = CC_{es}^{m} + IC_{es}^{m} + \sum_{i \in I} \sum_{n \in N} {EC_{n,i}}^{m} + {RC_{n,i}}^{m} - {SV_{n,i}}^{m} \tag{6.11}$$

$$CC_{es}^{m} = \begin{cases} c_{cc}\left(P_{dist,es}^{m} - P_{dist,es}^{m,max,past}\right) & P_{dist,es}^{m} > P_{dist,es}^{m,max,past} \\ 0 & P_{dist,es}^{m} \leq P_{dist,es}^{m,max,past} \end{cases} \tag{6.12}$$

$$IC_{es}^{m} = \begin{cases} c_{ic}({r_{cluster}}^{m} - {r_{cluster}}^{max,past}) & {r_{cluster}}^{m} > {r_{cluster}}^{max,past} \\ 0 & {r_{cluster}}^{m} \leq {r_{cluster}}^{max,past} \end{cases} \tag{6.13}$$

$${EC_{n,i}}^{m} = K_i(\theta_{n,i}^{m} - \theta_{n,i}^{m-1})^{\alpha_i} + {PEC_{n,i}}^{m} = K_i(\theta_{n,i}^{m,new})^{\alpha_i} + {PEC_{n,i}}^{m} \tag{6.14}$$

$${PEC_{n,i}}^{m} = c_{pe,i}\theta_{n,pe}^{m,new} = c_{pe,i}\, sf_{pe}\, \varepsilon_{pe,i}\, \theta_{n,i}^{m,new} \tag{6.15}$$

$${SV_{n,i}}^{m} = k_i^{salv} \frac{{L_i}^{\theta}}{{L_i}^{max}} K_i(\theta_{sv,n,i}^{m})^{\alpha_i} \tag{6.16}$$

6.2.1.7.2 <u>*Operational Costs*</u>

The operational costs C_{opex}^{m} detailed in (6.17)-(6.21) below account for expenses due to heating via a traditional coal stove C_{coal}^{m} , maintenance of fees associated with each energy system component C_{om}^{m} and the economic value of unmet electricity load C_{ul}^{m}. Expenses associated with space heating are found through the required monthly coal mass ${M_{coal}}^{m}$ and the local average price by weight c_{coal} – assumed invariant with time [247]. O&M costs are found as a function of total energy system installed capacity ${\theta_{i,es}}^{m}$ and respective periodic estimated maintenance fee $c_{om,i}$ for each component i, using recently reported average values for small scale systems [250]. To compute the economic value of energy not served, the total monthly unmet load $ED_{ul,es,final}^{m}$ for the entire system is multiplied by the curtailment fee c_{ul}, as shown in (6.20). It is assumed that a *5 % monthly shortage is allowed*, thus no charges are associated below that level. The value for c_{ul} is estimated through the Willingness to Pay (WTP) approach, which relates perceived marginal benefits to the maximal price a customer would be willing to pay for a service [251]. It is, therefore, assumed equivalent to the local cost of diesel which would be needed to supply the unmet load – neglecting generator investment cost and assuming standard conversion efficiencies. Cash flows resulting from grid interactions can be either positive or negative, based on the monthly energy balance P_{grid}^{m}, and are computed using either the reported discounted tariff c_{grid} or subsidized price received p_{grid}.

$$C_{opex}^{m} = C_{coal}^{m} + C_{om}^{m} + C_{ul}^{m} \tag{6.17}$$

$$C_{coal}^{m} = c_{coal}\, {M_{coal}}^{m} \tag{6.18}$$

$$C_{om}^{m} = \sum_{i \in I} c_{om,i} {\theta_{i,es}}^{m} \tag{6.19}$$

$$C_{ul}^{m} = c_{ul}\, ED_{ul,es,final}^{m} \tag{6.20}$$

$$C_{grid}^{m} = \begin{cases} c_{grid} P_{grid}^{m} & P_{grid}^{m} \geq 0 \\ p_{grid} P_{grid}^{m} & P_{grid}^{m} < 0 \end{cases} \tag{6.21}$$

6.2.1.8 System Mass

The total system mass M_{es}^{m} can be found from (6.22) for any month m as the sum of the weight of the installed renewables M_{res}^{m}, batteries M_{batt}^{m}, power electronics M_{pe}^{m}, and cabling length required for connection M_{cb}^{m}. To reflect mobility considerations associated with a migratory lifestyle, an upper limit on energy system weight $M_{es,max}^{m}$ may be imposed, captured by (6.23).

$$M_{es}^{m} = M_{res}^{m} + M_{batt}^{m} + M_{pe}^{m} + M_{cb}^{m} \tag{6.22}$$

$$M_{es,max}^{m} \leq M_{es,max}^{m} N_{n} \tag{6.23}$$

6.2.1.9 Resilience and Energy Security

The outputs from the power flow model introduced in 6.2.1.5 are then used to estimate the resilience and energy security of the mobile micro-grid system. While varied measures are available depending on the application context[37], total unmet load (TUL) is well justified in an energy access context. This metric TUL^{m} is computed monthly to capture seasonal differences, and maintain consistency with the rest of the model presented in this section. It is calculated as shown in (6.24), by summing the hourly unmet electricity load $P_{uel,es,ppp}^{h}$ resulting from system operation over 30 days (i.e. $H = 720\ hours$).

$$TUL^{m} = \sum_{h=0}^{H} P_{uel,es,ppp}^{h} \tag{6.24}$$

6.2.1.10 Energy System Model Summary

Figure 6.3 captures the interaction among different model components detailed in the former sections, leading to calculations of total monthly net system cash flow ($NCF^{t} = C_{es}^{m}$), emissions ${CO_2}_{eq.emissions}^{m}$, and unmet electricity load TUL^{m} for a single scenario. Note that energy security assessment in 6.2.1.9 only considers electricity needs, given all heating demand must always be met for nomadic communities living in extreme climatic conditions. Consequently, the budget and mobility constraints (formulated in (6.10) and (6.23)) are not applied when the monthly energy system

expenditure (defined in (6.9)) is based solely on coal. The total unconstrained cost to supply the necessary thermal energy is calculated through (6.18) based on (6.7) in this case.

Changes or upgrades to the energy system configuration are determined by the DRL agent policy captured by the green box, further detailed in 6.3. This policy or flexibility strategy selects capacity adjustment decisions over time based on progressive observations of uncertain variables in its state vector. As shown in Figure 6.3, action selection impacts the costs, carbon footprint, and energy security recorded for the flexible design in all remaining simulation steps. For the baseline rigid systems, on the other hand, these calculations are based purely on uncertainty realizations.

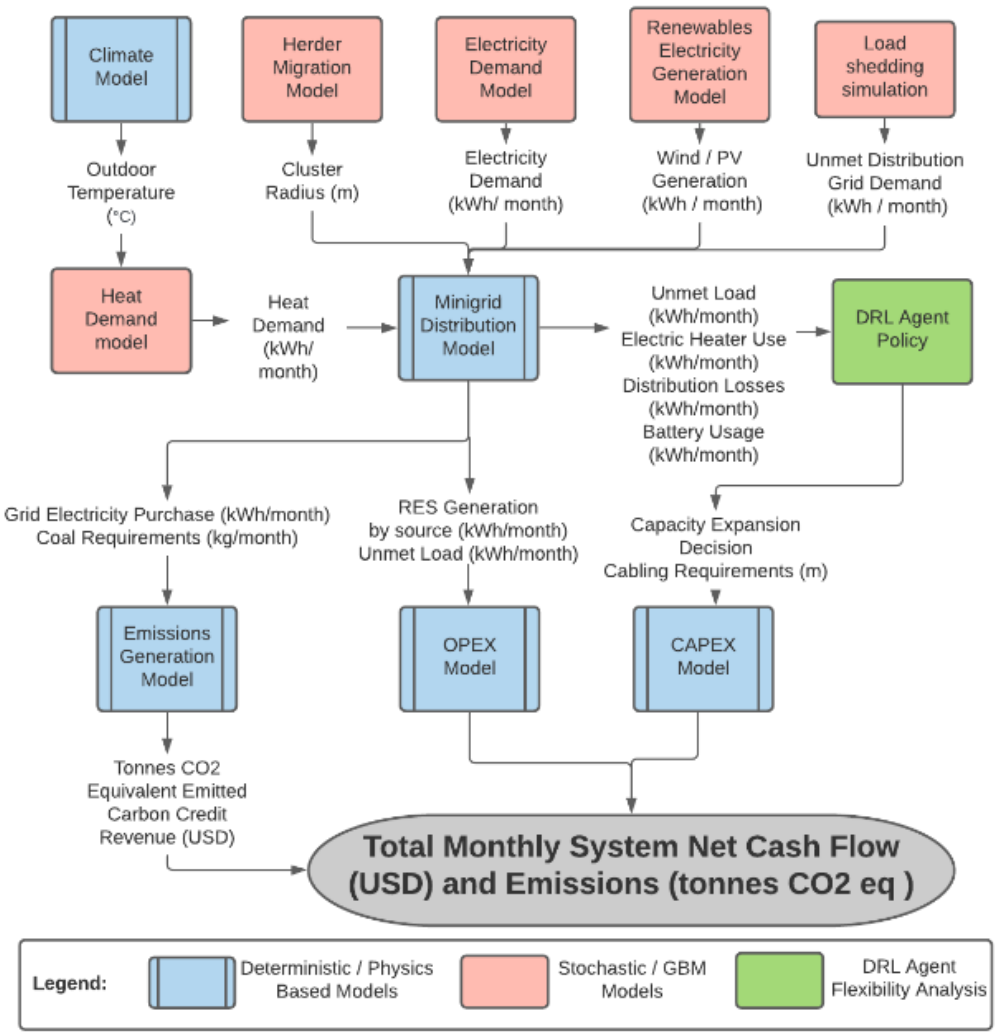


Figure 6.3: Model flowchart showing complex interactions among system components and simulations.

6.2.2 Background and Previous Work

Nomadic energy access and mobile energy systems are highly relevant in Mongolia, where it is estimated that 25-40 % of the national population continues to live a migratory lifestyle, normally moving between 4 and 30 times a year [252]. Changing climatic conditions, however, have increased the frequency of "zhuds" – extremely cold winters followed by extremely dry summers – making the traditional way of life more difficult for nomadic communities. This has resulted in an explosion of population migration to UB, where out of the approximately 1.4 million residents around 60% live in "ger areas" [252]. A "ger" refers to the traditional Mongolian herder dwelling, comprised of a basic wooden frame and a number of insulating layers to allow for mobility. On the other hand, "ger areas" represent the urban regions surrounding UB where the majority of these low-income communities now settle in the winter [253], and normally migrate to rural surroundings during the summer months, with annual movement ranges of around 7-8km [252]. Previous studies estimate nomadic community clusters are typically separated by 3.8km in the spring, 2.0km in the summer and 2.3km away in the autumn [254]. Figure 6.4 below graphically captures the migration patterns investigated, where MS1 represents operation in the area between UB and Hently in the summer, while MS2 is defined by migration to ger areas in UB during winter months, as in 6.2.1.1.

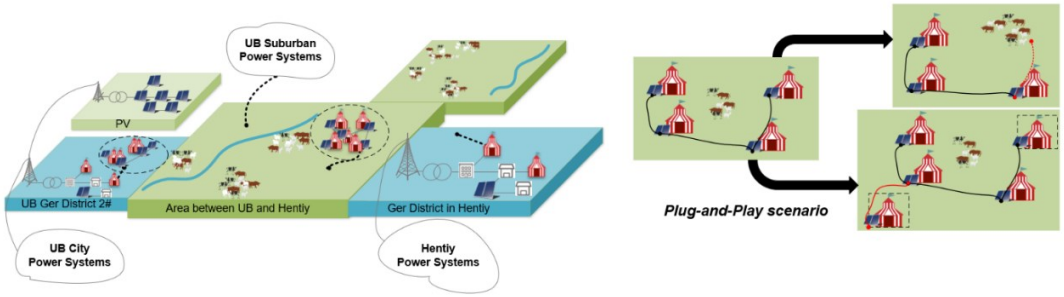


Figure 6.4: Spatial boundaries for typical migration pattern (left) and example multi-ger PP configurations (right)

6.2.2.1 Electrification

The national network infrastructure is largely unsuitable to provide reliable and sustainable electrification across the potential MS encountered. It currently relies primarily on coal (71%) and oil (25%) for centralised generation, with a large associated carbon footprint [205]. More specifically, during MS1 the Northern Energy System is the primary option for herder electricity connection [252]. Highly underdeveloped and unstable it mainly relies on imports from Russia [202]. It also only covers a fraction of the potential migration range, with 45% of rural gers unable to access electricity services consistently [200]. During MS2, on the other hand, gers are normally able to connect to the distribution grid in UB. This network, however, also experiences very frequent load shedding and blackouts, leading to limited energy security for ger communities [255].

In light of the above, household or community-based decentralised power supply systems are likely the most cost-effective option for achieving SDG7 in Mongolia [204]. In fact, more than 100,000 Small

Home Solar (SHS) systems and 5,000 small wind systems (SWS) have already been distributed through the Renewable Energy and Rural Electricity Access Project (REAP), with capacities typically ranging from 20 to 100W per ger [202,204]. REAP, however, was primarily based on capital investment subsidies and very low interest financing, thus not considering load evolution, mobility, or local resource assessment in capacity planning. Most of the systems currently installed are, as a result, either no longer in use or significantly undersized [204].

6.2.2.2 Heating Supply Decarbonisation

Electrification of nomadic communities has a very strong link with Mongolian government heating supply decarbonisation objectives, as residents of ger areas normally are unable to connect to the district heating network [247]. With the majority of the country in the extreme cold zone, there is substantial demand for space heating over a long winter season. As a result, it is estimated that ger areas contribute about 45% to 75% of UB's annual average fine particulate matter ($PM_{2.5}$) emissions [247,256]. This is driven by the inefficient combustion of raw coal, biomass, or household waste, the dominant heating supply sources for low-income gers, making UB one of the worlds most polluted cities in winter months, with significant associated health issues [257]. Based on a median income of only around 350 USD ger^{-1} $month^{-1}$, costs for space heating also contribute around 15-20% of annual ger expenditures [247,258]. During winter, up to half of their limited monthly income may be spent on meeting heating demand alone, prioritising the cheapest source available.

Given the severity of and urgency of the air pollution issue, most previous works on ger energy systems seek to address decarbonisation of the heating supply under the extreme climatic conditions typically experienced [201,247,253]. Traditional coal-stoves are used in most gers, thus these studies normally focus on either benefits from using improved stove designs [259,260] and/or better thermal insulation [201,247]. Partly, this is because these works are centred on ger areas in the outskirts of UB where the grid capacity is a limiting factor to alternative electricity-based heating solutions, and the pollution is particularly severe. Most recently, however, [261] presented the experimental investigation and simulation of a solar PV generator combined with an EH for a single ger in UB, as compared to one using a traditional coal stove. The results from the study over a single year suggest a transition to electricity-based heating supply system can lead to improved indoor environmental quality, cost reductions, and significantly lower heat energy consumption.

6.2.2.3 Local Context Constraints

Given the mean monthly coal expenditure reported by [247], the total energy systems max expenditure $C^{m}_{es,max,ger}$ is limited to USD 110 $month^{-1}$ ger^{-1}. This results in approximately USD 20-60 $month^{-1}$ ger^{-1} available for capacity expansion, depending on the respective coal expenditure, formulated at the entire 18-ger energy system level. Furthermore, it is well established that a typical 5-wall Mongolian

ger weighs 250-400 kg, with an additional 75-100 kg for a standard coal stove and at least one 20kg bag of raw coal normally used during migration [201,253]. Considering the above, an upper limit on total energy system weight $M^{m}_{es,max,ger}$ of 120 kg ger^{-1} is preliminarily estimated.

6.2.3 Herder Migration

The assumptions used to model ger migration patterns and spatial dispersion uncertainty were developed and validated through direct engagement with a team of local consultants [190] and an interactive workshop with regional energy sector stakeholders [199]. Consequently, it was identified that 18 gers is representative of a typical ger nomadic community. Recent studies suggest that the average area occupied by such a community over the migration season is around 147 km^2, leading to an average seasonal cluster radius of **6.85 km** at project start [190,254]**.** Furthermore, previous works suggest climate change is likely to lead to growth in overall migration distance and spatial dispersion over the 30-year horizon [252]. With the aim to capture how this uncertainty may influence optimal energy system configurations, mean annual growth rate and volatility are preliminarily estimated based on available data. The annual cluster radius $r_{cluster}{}^{y}$ evolutions are thus simulated according to (3.11)-(3.12). The resulting potentially large differences between projected (dashed line) and realised (solid lines) migration radius are captured by Figure 6.5, showing the impact of stochasticity on both long-term evolution and intra-year variability for a few select example scenarios.

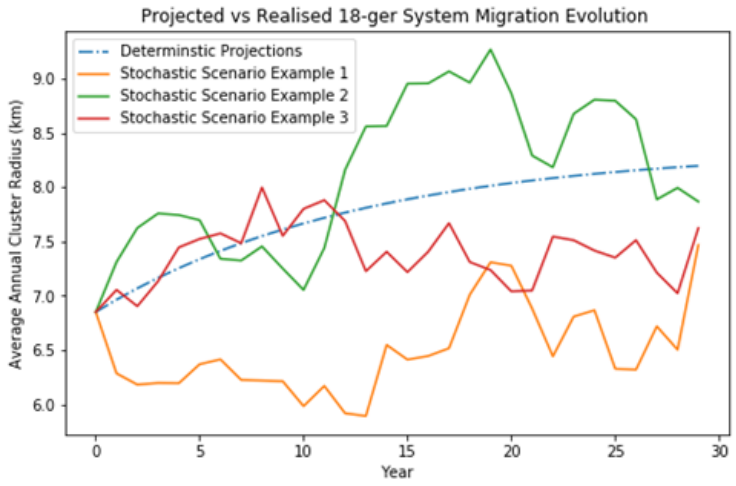


Figure 6.5: Deterministic projections **(dashed line)** and 3 randomly selected example stochastic simulated scenarios **(solid lines),** showing the average annual (over the migration season) 18-ger cluster radius in km (y-axis), over the 30-year project horizon (x-axis). The projected migration growth is assumed to be driven by the climate change effects, as shown by the deterministic projections **(dashed line),** although it is recognized the magnitude of this impact is highly uncertain both in its long-term evolution and intra-year variability, as captured by the different stochastic example scenarios **(solid lines).**

6.2.4 Electricity Demand (ED)

The baseline daily electricity load profile for a typical ger household was estimated through interactions with a local consulting team [190], given the limited literature data available. These values

suggest most nomadic households continue to follow a more traditional way of life for now, typically only consuming enough electricity to operate a 20W lamp and 80W refrigerator. In winter months, however, the local climatic conditions as well as economic constraints mean that the refrigerator is not needed nor used. This assumption is maintained throughout project horizon, with starting electricity demand estimates of roughly 63 and 6 kWh month^{-1} ger^{-1} during the summer and winter months, respectively. Nonetheless, this load is likely to grow significantly over the next 30 years, with recent ones experiencing a 5-7% annual increase in electricity demand, driven by a combination of more productive uses of electricity integrated into nomadic lifestyles, higher indoor environmental standards, and the adoption of low-cost electric appliances [202,206,253].

Baseline load profile projections are thus formulated based on the integration of a 213W TV with a 30W antenna, more light bulbs, and a number of additional electric appliances (i.e., radios, cell-phone) by years 15 and 30, respectively [190]. The stochastic electricity demand model is defined from (3.11)-(3.12). The baseline daily load profiles for summer days are shown on the left in Figure 6.6 below at the starting point (Year 0), midpoint (Year 15) and end (Year 30) of the project. Long term evolutions and variation under uncertainty for different seasons, on the other hand, are captured on the right in Figure 6.6. For power flow calculations in 6.2.1.5, it is assumed that these same hourly load profile curves are maintained, the magnitude of which is scaled proportionally to the actual realized monthly demand as compared to the one projected at time of writing.

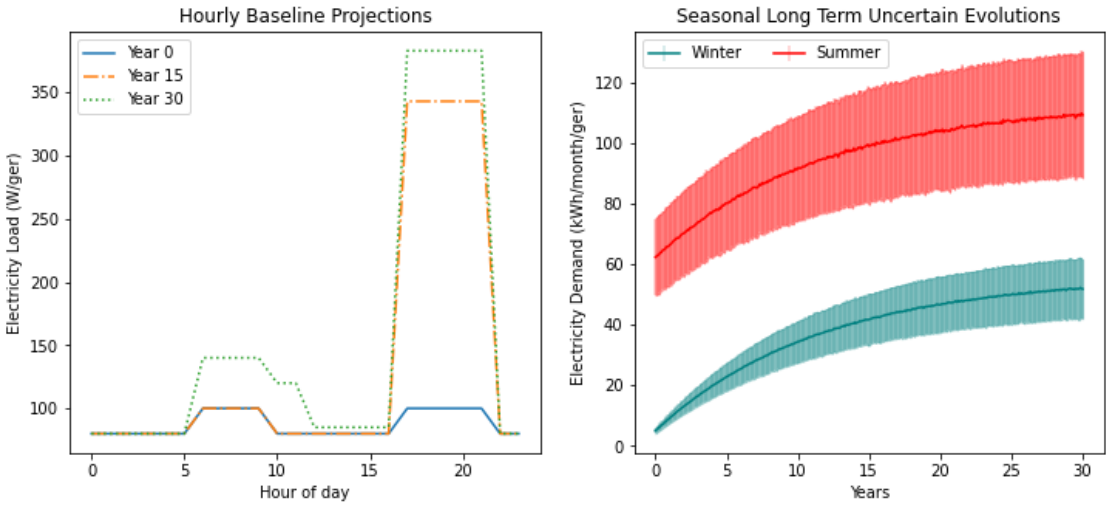


Figure 6.6: Hourly projected load (left) and seasonal long term stochastic evolutions of electricity demand (right).

6.2.5 Heating Demand (HD)

As a baseline for estimating heating demand (HD), it is assumed that all 18 gers in the systems evaluated have 5 walls with standard insulation. While the benefits of improved insulation for gers have been discussed in several other studies [201,258,261], the standard insulation case defined in these works is still representative of the majority of ger dwellings. Seasonal demand calculations are

obtained through consultation with a local team [190] and are aligned with values found in the literature [201,247,258,261]. Respective demand-temperature correlations are then used to extrapolate monthly profiles based on the assumptions given in 11.6, resulting in starting demand of 11,500 kWh/ year per ger. Volatility for heating demand is assumed to be approximately equivalent to that on year-to-year climatic conditions, found via standard regression analysis of the average *winter* temperatures between 1901-2020 [194]. The stochastic HD model is thus formulated yearly according to (3.11)-(3.12), including potential climate change induced growth [193,198,262]. The impact of this evolution and differences in potential Year 30 HD profiles is shown for 3 stochastic scenarios in Figure 6.7. As for ED above, the seasonal profile *distribution* is maintained across years simulated, and the value in each timestep scaled proportionally to the one at project start [190].

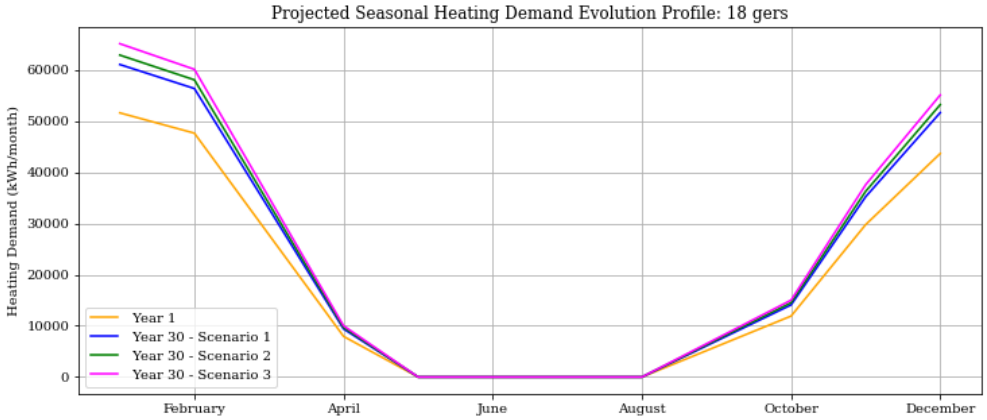


Figure 6.7: Projected starting and end of project life heating demand monthly profiles for 3 scenarios.

6.2.6 Renewable Energy Generation

For the system sizes and representative locations considered in this manuscript, a mean annual capacity factor (CF) of 19.8% for solar PV is obtained, similar as reported in [261]. The initial estimate for annual wind capacity factor is slightly higher at 22.2%. This is lower than values from other wind generation sites in Mongolia, as the small-scale systems selected are operating at reduced wind speeds and conversion efficiencies compared to utility level ones [204,206]. Still, both CF estimates are consistent with those which could be extrapolated from alternative data sources [263,264] or with previous studies [195,205,206].

Please refer to 11.6 for the complete list, values and ranges of the parameters used in this study.

6.3 DRL Implementation

6.3.1 DRL Flexibility Analysis Formulation

The DRL formulation for the mobile micro-grid design and planning problem presented in this chapter is characterized by the general objective and problem statement captured in (3.10). The only key difference, however, is that due to financial limitations for the targeted communities the minimisation of expected net present cost (ENPC) is targeted here – as opposed to ENPV

maximization in the other chapters. The remaining procedure with the novel proposed approach, nonetheless, remains the same, with the environment developed through the OpenAI gym framework[161]. All the steps involved in DRL Flexibility analysis are further detailed below.

1. **Environment Formulation:**
    a. State Space: In this context, the engineering system evolution strategy is primarily dependent on the periodic demand for different energy services, installed generating capacities and migratory patterns, among other factors. With the aim of reflecting this, the states for the agent are defined through a box observation set according to (6.25) below. where ED^t and HD^t represent the electricity and heating stochastic demand realizations in timestep t, respectively. $r^t_{cluster}$ and ${L_{cb}}^t$ measure the uncertain spatial dispersion of the mobile system investigated and are used in cabling investment and distribution loss calculations. As in the former section, ${\theta_i}^t$ represents the installed *nominal* capacity for each system component i considered. $ED^t_{ul,es,final}$ is included to stimulate the explicit consideration of unmet load (i.e. approximation of resilience), outside of charges incurred, while M^t_{es} is used to account for increases in system mass as it evolves over time. Budget constraints are implicitly evaluated based on recorded rewards. All state variables are normalized based on the maximum and minimum values obtained for them over 10,000 simulated scenarios, independent of the environment. Time limit awareness is also implemented as previously discussed.

$$s^t = \left[ED^t, HD^t, r^t_{cluster}, {L_{cb}}^t, {\theta_{pv}}^t, {\theta_{wind}}^t, {\theta_{eh}}^t, {\theta_{bess}}^t, ED^t_{ul,es,final}, M^t_{es}, T-t\right] \quad (6.25)$$

    b. Reward Function: Based on the energy access objectives of this work, the focus is on minimisation of cost needed to provide electricity and heating services to the targeted communities. Note that a reward penalty is incurred if local budgetary (6.10) or mobility constraints (6.23) are not met in a particular timestep, as shown in (6.26). The same penalty magnitude is applied for both constraints as they were both deemed equally important in the targeted context based on initial stakeholder interactions, although this may be easily modified if more information on decision-makers priorities became available. This allows the DRL agent to implicitly learn their form, looking for system solutions which are able to minimise the occurrence of constraint violation. While more advanced approaches are available to ensure constraints are robustly met, particularly in safety critical applications [174,221,265], they are typically associated with a high computational expense and are complicated to implement. As such, given in this setting the constraints represent socio-cultural considerations rather than safety issues, the simpler constraint violation penalty method is deemed

justified. Furthermore, this helps reduce the computational expense on an already very highly dimensional problem.

$$r^t = \begin{cases} C_{es}^m = C_{inv}^m + C_{opex}^m + C_{grid}^m + C_{CO2}^m & if \ (6.10) \text{ and } (6.23) \textbf{ are True} \\ -USD\ 1{,}000 & otherwise \end{cases} \quad (6.26)$$

2. **Uncertainty Quantification**: The characterization of uncertainty considered in this design and planning problem was already extensively discussed in 6.2 and the interaction among different modelling components is graphically captured by Figure 6.3. It is again remarked in this section that the limited data available lead to a very significant uncertainty space, however all parameters and ranges for stochastic simulations were developed in collaboration with local stakeholders[190,199]. This leads to very important differences among simulated scenarios, highlighting the need for a data-driven approach to analyse engineering system flexibility in what would otherwise quickly become a computationally intractable optimization using standard methods.
3. **Action Space Creation:** The decision-making possibilities for the agent are also selected with the aim of reflecting the ones available to a ger community, through a multi-discrete action set ($\in A_{i,j}$). Through this formulation, the DRL decision space is such that, at any given time step, the agent may decide to leave the energy system configuration unchanged. On the other hand, based on estimated value function and policy, a decision enabled by flexibility may be made to alter the state of the microgrid. In that case, the multi-discrete set I_j means a decision is made on which technology i to act on, and whether to pursue capacity abandonment, expand by 500W or expand by the (rounded) maximum number of modules θ_{Budget}^t possible under the budgetary constraints C_{Budget}^t for that period. As actions are selected at the entire system level, a simple heuristic is implemented to determine which node to implement the capacity adaptability decision on. For expansion decisions on solar PV, wind or BESS, additional component capacity is assumed to be installed at the node with the largest unmet load in the previous month $ED_{uel,max}^{n,t-1}$. Abandonment actions or EH expansions, on the contrary, are allocated to the node with the largest *excess* generation $P_{res,excess}^{n,t-1}$ prior to implementing the flexibility option. The full action space is mathematically captured by (6.27), where $I_j = [-500W, 500W, \theta_{Budget}^t]$ and only one capacity adjustment action available per monthly decision-making timestep.

$$a^t = \begin{cases} 0 \rightarrow & No\ capacity\ adjustment \\ I_j \rightarrow & Adjust\ technology\ i\ capacity\ by\ j \end{cases} \quad \text{for t} \in \text{T}, i \in I \quad (6.27)$$

Figure 6.8 below captures the internal evaluation process, going from state observation (left) to action selection (right) using a DNN parametrized agent policy, where components are evaluated within the environment technological boundaries. The iterative interaction feedback loop guiding policy updates during training, is visualized through the changes in system state, reward and policy driven by a capacity adjustment decision made at time t on $t+1$.

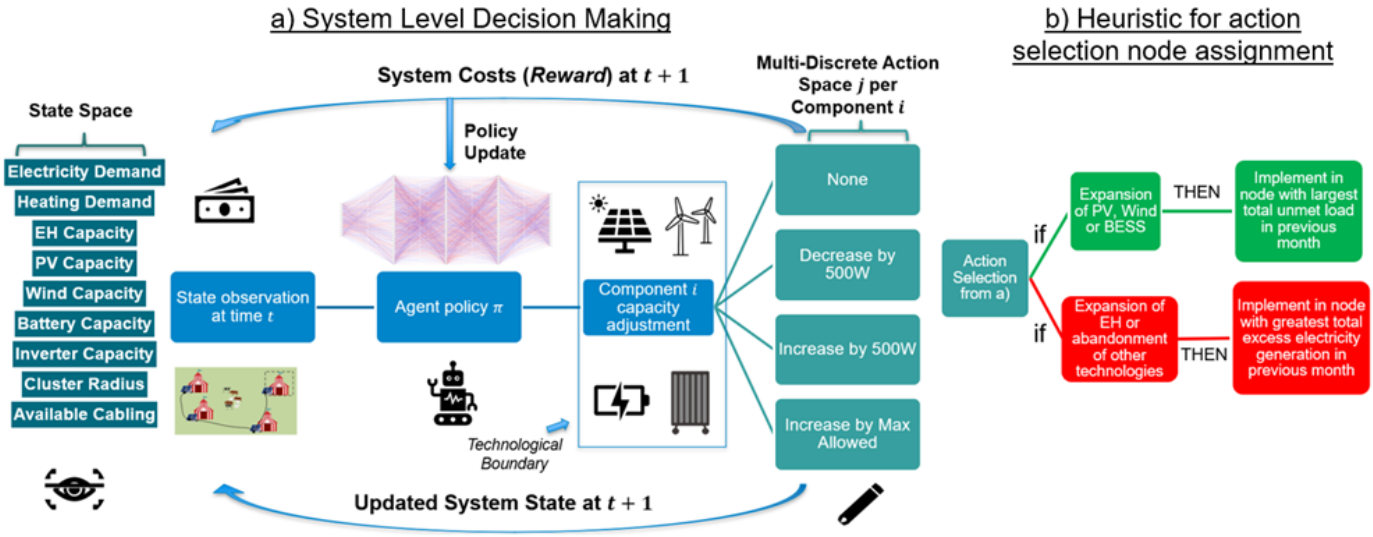


Figure 6.8: a) Graphical overview of DRL system level decision making process going from microgrid state space input (left) to action selection output (right) b) Simple heuristic used for allocation of system level DRL actions to individual microgrid nodes, either the ones with either largest unmet load in previous month (top) or greatest total excess electricity generation (bottom).

6.3.2 Algorithmic Approach

Considering the dimensionality and computational expense of the problem tackled, the ACER (*Sample Efficient Actor Critic with Prioritized Experience Replay)* algorithm is implemented [266]. Each component is captured by a separate neural network, allowing different strategies to be implemented in policy as compared to value function updates. Experience replay allows agents to focus on the interactions most significant to experienced reward, improving convergence rate during training [141]. Sample efficiency was motivated by the complexity of the model, evaluated with different timestep resolutions for operation and decision making, meaning each environment episode simulation is relatively expensive computationally. This is achieved in ACER by combining several ideas from previous works with novel developments.

More specifically, multi-step impacts on state-action value function $Q_{ret}(s^t, a^t)$ approximations for each generated trajectory are found recursively using the Retrace algorithm [267], as shown in (6.28). Multi-step returns used allow for significant reductions in policy gradient bias and learning iterations needed for the Critic component of the algorithm. A duelling architecture is implemented, through

which the vectorized value function estimates $Q_{\emptyset}(s^{t+1}, a^{t+1})$ as well as the policy $\pi_{\emptyset}(a^t|\, s^t)$, are outputs of the DNN with parameters $\emptyset$. As shown in (2.2), the estimate $V_{\emptyset}(s^{t+1})$ is then easily found as the expectation of $Q_{\emptyset}$ under $\pi_{\emptyset}$. Additionally, note that γ is mathematically equivalent to the financial discount factor $\frac{1}{1+\lambda}$ shown in the generic problem formulated in (3.10). It is used to in DRL to account for the different value of immediate rewards compared to long term ones. The truncated importance weight $\bar{\rho}^{\,t}$ under current behaviour policy μ, is thus computed as shown in (6.29)[152].

$$Q_{ret}(s^t, a^t) = R^t + \gamma\, \bar{\rho}^{\,t+1}[Q_{ret}(s^{t+1}, a^{t+1}) - Q_{\emptyset}(s^{t+1}, a^{t+1})] + \gamma V_{\emptyset}(s^{t+1}) \tag{6.28}$$

$$\bar{\rho}^{\,t} = \frac{\pi_{\emptyset}(a^t|\, s^t)}{\mu(a^t|\, s^t)} \tag{6.29}$$

Policy updates are then conducted in a very similar manner to Trust-Region Policy Optimization [149], although using the running average policy network as a baseline instead. The policy is thus decomposed into a probability distribution f, and network capturing the properties of the policy distribution $\tau_{\emptyset}$, with the update split into two stages. In the first stage, a linearized Kullback–Leibler (KL) divergence constraint is combined with a standard loss minimization optimization. In the second stage, backpropagation is used to compute the derivatives with respect to the policy parameters, adopting a trust region based on the distribution f and advantage function. The gradient of ACER is thus defined as shown by (6.30) below, where c is the correction term coefficient for cases with very high variance.

$$\begin{aligned}\hat{g}_{ACER}{}^{t} = {} & \bar{\rho}^{\,t} \nabla_{\tau_{\emptyset(s^t)}} \log f(a^t|\, \tau_{\emptyset}(\, s^t)\, [Q_{ret}(s^t, a^t) - V_{\emptyset}(s^t)] \\ & + \mathbb{E}_{a\sim\pi}\left(\left[\frac{\bar{\rho}^{\,t}(a) - c}{\bar{\rho}^{\,t}(a)} \right] \nabla_{\tau_{\emptyset(s^t)}} \log f(a^t|\, \tau_{\emptyset}(\, s^t)[Q_{\emptyset}(s^t, a^t) - V_{\emptyset}(s^t)] \right)\end{aligned} \tag{6.30}$$

The full pseudocode for the ACER agent policy optimisation problem as applied to the design and planning of mobile micro-grids under uncertainty is given by Algorithm 3 below. Generally, Algorithm 3a is first used to generate potential trajectories and perform updated in an on-policy manner. Several replay steps are then conducted by calling ACER off-policy. Please refer to the original publication for further details and definitions [268].

Algorithm 3: ACER applied to mobile micro-grid design and planning, adapted from [268].	
Algorithm 3a: Master algorithm ACER for discrete action space.	
1:	Assume global shared parameter vectors θ and θ_v Assume ratio of replay r

2:	**Repeat**
3:	Call ACER <u>on-policy</u> according to Algorithm 3b
4:	$n \leftarrow Poission\ (r)$
5:	**for** $i \in [1, \dots., n]$ **do**
6:	Call ACER <u>off-policy</u> according to Algorithm 3b
7:	**end for**
8:	**Until** Max training iterations or computational budget reached.
Algorithm 3b: ACER policy optimisation with discrete action space.	
1:	Initialize parameters $\theta' \leftarrow \theta$ and $\theta'_v \leftarrow \theta_v$
	Reset gradients $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$
2:	**If not** On-Policy (see Algorithm 3a) **then**:
3:	Sample the trajectory $[s^0, a^0, r^0,\ \mu(\cdot\,|s^0), \dots., s^T, a^T, r^T,\ \mu(\cdot\,|s^T)]$ from replay memory
4:	**Else:**
5:	Initialize starting energy system state s^0 based on (6.25)
6:	**End if**
7:	**for** $t \in [1, \dots., T]$ **do:**
8:	Compute policy probability distributions $f(\cdot\,|\emptyset_{\theta'}\ (s^t))$, $Q_{\theta'}(s^t, \cdot)$ and $f(\cdot\,|\emptyset_{\alpha}\ (s^t))$
9:	**If** On-Policy (see Algorithm 3a) **then**:
10:	Adjust energy system capacity by performing a^t from (6.25) according to $f(\cdot\,|\emptyset_{\theta'}\ (s^t))$
11:	Receive reward r^t from (6.26) and update energy system state s^{t+1} based on Figure 6.8
12:	Update action probability distribution $\mu(\cdot\,|s^t) \leftarrow f(\cdot\,|\emptyset_{\alpha}\ (s^t))$
13:	**End if**
14:	Update truncated importance weight $\bar{\rho}^t$ based on *minimum* of [1, value from (6.29)]
15:	**End for**
16:	Update multi-step impacts on action-state value function $Q_{ret}(s^t\ , a^t)$ according to (6.28), set to 0 for terminal states (i.e., $t = T = 30\ years = 360\ months$)
17:	**for** $t \in [T-1, \dots.,0]$ **do:**
18:	$Q_{ret} \leftarrow\ r^t + \gamma Q_{ret}$ using updated multi step impacts from line 16
19:	Update value function based on action selections over episode using $Q_{\theta'}(s^t, a)$ and $f(a|\emptyset_{\theta'}\ (s^t))$
20:	Compute quantities g and k needed for trust region updating
21:	Accumulate gradients w.r.t. θ' according to (6.30)
22:	Accumulate gradients w.r.t. θ'_v as: $d\theta_v \leftarrow d\theta_v + \nabla_{\theta'_v}(Q_{ret} - Q_{\theta'_v}(s^t, a))^{\,2}$
22:	Update retrace target Q_{ret} for next policy improvement iteration
24:	**End for**
25:	Perform asynchronous update of θ using $d\theta$ and θ_v using $d\theta_v$
26:	

27:	Update the average policy network : $\theta_a \leftarrow \alpha\theta_a + (1-\alpha)\theta$

6.3.3 Training and Testing

The ACER agent described in the former section is trained using a budget of 72 million monthly decision timesteps, equivalent to 200,000 simulated full 30-year episodes. As performed on a standard laptop Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz, the complete training process takes approximately 8-9 hours. All training scenarios are stochastically generated to be *i.i.d* from each other, thus leading to an *infinite* amount of potential scenario evolutions and state-action pairs. This was motivated by the need to develop a more generalizable adaptability strategy under different realizations of uncertainty. No exactly equal trajectories are thus possible inside the environment, and training process is fully non-anticipative, with all operational decisions based purely on recorded observations at a particular timestep. Hyperparameters are found through Optuna across 100 experiments[170], with final training conducted using 50 environment steps per policy update, parallelized through 4 processes, a correction term of 12.3 and a replay ratio of 5. The replay ratio and denotes the average relative replay learning per policy learning – which only kicks in after 20,000 replay starts in this case. Please refer to 11.7 for the full list of hyperparameters used.

The DNN architectures implemented for both the actor and critic component of Algorithm 3 are fully connected, with input layers matching the problem feature space described above. The output layer for the actor is composed of twelve neurons, representing the multi-discrete action selection process depicted in Figure 6.8, whereby four independent SoftMax distributional operators are then used to characterize decisions for each component. Additionally, five hidden layers with 64 neurons each are implemented to approximate the complex environment stochastic functions. Crucially, it should be noted that this approximation derived from the DNN captured by (3.10) does not always guarantee the optimality of the results obtained - especially when considering the non-convexity of the function evaluated. Nonetheless, several studies suggest strong convergence properties for DRL to low lying local minima when using gradient-based techniques, which employ a number of heuristics to reduce the likelihood of obtaining sub-optimal solutions [269-276]. The Stochastic Gradient Descent Momentum method [277] is thus implemented for DNN training parameter updates, using the exponentially moving average of the trust-region adjusted values through (6.30).

The complete learning curve for the ACER agent on the case study detailed in this chapter is given by Figure 6.9 below, averaging results from 5 separate training runs using different random initialization seeds. Note that only the best training run is used for relative performance analysis in 1.1. The training reward evolution recorded confirms the proposed approach to be effective at analysing flexibility even in such a highly dimensional and uncertain engineering system. The bulk of agent learning seems to be focused on the first quarter of the training process (i.e., 0 to 18M timesteps),

where the ENPC of the mobile microgrid for Mongolian nomadic communities is reduced by almost 50%. Past this point, the policy begins to gradually converge, as the budgetary and mobility constraints have been learned through the reward function. The final quarter of the training process (i.e., 54M to 72M timesteps) suggest the development of a near-optimal policy as the objective function only marginally improves, with more limited variation among training runs.

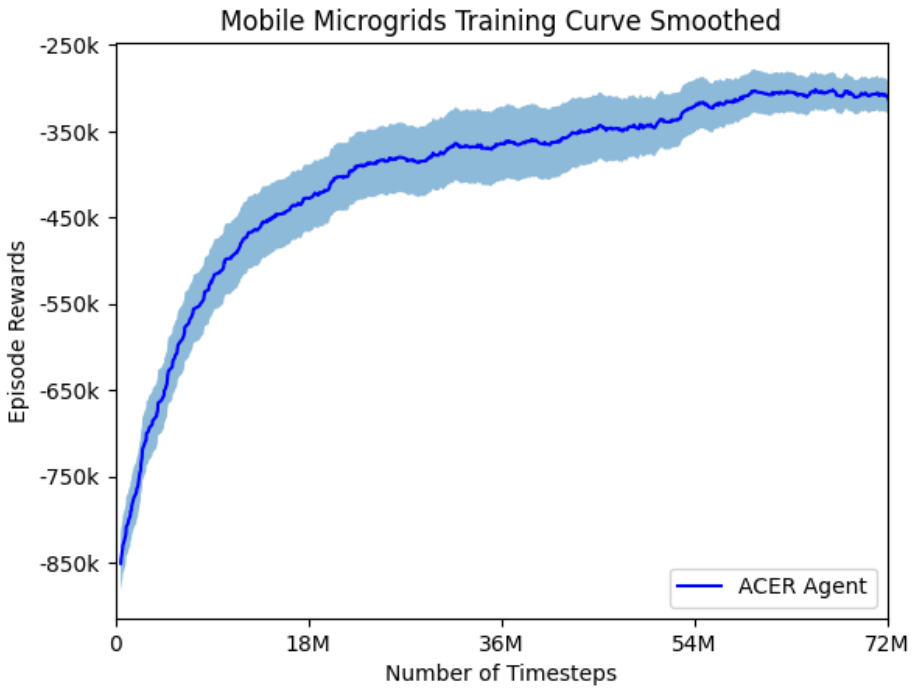


Figure 6.9: Training episode reward evolution for ACER agent on Mongolian mobile microgrids environment. The y-axis estimated the total energy system cost per episode while the x-axis shows the number of *monthly* decision timesteps implemented for training. The solid line shows smoothed average (using a window size of 50 episodes) among 5 runs. The shaded area the standard deviation recorded across the same training runs.

The reward penalty from (6.26) is then removed in the testing environment, allowing evaluation of the true ENPC rather than the one obtained through the modified training reward function. Furthermore, to evaluate flexible system performance as compared to the baseline designs, all results are reported using the same 2,000 simulated *out of sample* episodes. This means that while trained on the same stochastic model *parameters*, no complete scenario used in testing would have been encountered during training. The different time step resolutions implemented are summarised by Figure 6.10 below.



Figure 6.10: Summary of time-step resolutions implemented for different levels of the mobile micro-grid simulations used for the proposed DRL approach in this chapter.

6.4 Relative Performance Analysis

6.4.1 Summary

The energy system design alternatives discussed in the remainder of this section are summarised in Table 8 below, aiming to capture the benefits of flexibility and a systems approach to planning for ger communities. Note that, as in Chapters 4 and 5, the DE algorithm is applied to the optimisation of all baseline design (BD) alternatives, using the inflexible capacity planning objective function introduced in 3.4.1. The stochastically optimal rigid baselines results suggest the deployment of a SHS system for BD1 (θ_{pv}^{BD1} = 525 W/ger). This is approximately 5-10 times with what is currently available to herders, or roughly twice that obtained using commercial software such as HOMERPRO, which neglects load growth. Interestingly, the optimal PV capacity in BD2 is equal to that found in BD1, although with a roughly matching EH capacity (625 W/ger) for thermal energy supply. In BD2, therefore, the EH units are run almost exclusively with power purchased from the UB grid. BD3 and BD4 present very similar initial design capacities to their counterparts which are optimised for no PP connection (BD1 and BD2), albeit with reductions in BESS requirements of roughly 8% and 10%, respectively. System evolution for FD, on the other hand, varies based on scenarios, although with final capacities generally much greater (and diversified) than those obtained via rigid baselines.

Table 8: Summary of nomadic ger energy system design alternatives evaluated in this study.

Ger Energy System Alternative	**Description**
Flexible Design (FD)	18-ger multi-energy PP *microgrid* design and adaptability strategy
Flexible Nanogrid Design (FND)	DRL based single nanogrid unit design and adaptability strategy

Baseline Design 1 (BD1)	Single nanogrid unit stochastically optimal rigid design with no EH
Baseline Design 2 (BD2)	Single nanogrid unit stochastically optimal rigid design with EH option
Baseline Design 3 (BD3)	18-ger microgrid stochastically optimal rigid design with no EH
Baseline Design 4 (BD4)	18-ger microgrid stochastically optimal rigid design with EH option

The combined relative *out of sample* performance of the 18-ger energy system design alternatives optimised using the model detailed in 6.2.1 are summarised by below. They show that the FD yields improvements across all metrics estimating economic feasibility, sustainability, and energy security of alternatives, highlighting the importance of embedding capacity expansion flexibility under uncertainty. Note that, as presented in 3.4.5, VoF refers to improvements achieved by FD for each metric as compared to the *best performing benchmark design on the same metric*.. These results suggest that FD may be highly valuable for ger communities' electrification along different risk and utility metrics, as discussed in more detail below.

Table 9: Design decision making table for alternatives evaluated in this study over 2000 out of sample scenarios.

Metric	**Objective**	**BD1**	**BD2**	**BD3**	**BD4**	**FND**	**FD**	**VoF**
ENPC (USD)	*Economic Feasibility*	439,212	413,771	428,088	386,459	373,503	328,178	**85,593**
VaR, 5% (USD)	*Economic Risk*	492,314	516,332	480,058	473,062	419,512	367,421	**124,893**
VaG, 95% (USD)	*Economic Upside*	381,456	370,591	365,221	352,669	338,229	293,404	**77,187**
Lifetime Emissions (t. CO_2 Eq.)	*Sustainability*	10,133	14,169	9,808	13,375	4,937	4,601	**5,532**
Expected TUL (kWh)	*Energy Security*	29,992	29,992	28,102	27,171	8,274	7,188	**22,804**

6.4.2 Expected Energy System Cost

The low-income communities targeted by these studies motivate prioritising cost as an initial objective. FD achieves significant improvements in this front particularly as compared to BD1 and

BD2, reducing ENPC by 25.3% and 20.7%, respectively. The PP enabled inflexible solutions BD3 and BD4 perform slightly better due to the benefits of interconnection, although FD still results in an ENPC 23.3% lower than BD3 and a 15.1% reduction from BD4. Expected cost reductions compared to HOMER deterministic design are even more striking at around 41%, although highly dependent on the economic value of energy not served c_{ul} estimated earlier. Reduction in energy system cost in any given scenario are achieved by FD through an improved adaptability strategy under uncertainty. More practically, this results in decreased expenditure on coal over time, reduced charges due to unmet load, and discounting of capital costs through modular deployment of components, as shown in Figure 6.14. The greatest economic VoF is recorded for VaR, with FD achieving over a 30% reduction in downside costs as compared to alternatives investigated, although significant improvements are also found in terms of upside potential (VaG). Furthermore, as the capacity expansion strategy developed is highly dynamic and dependant on different realisations of uncertain variables, FD performs much more consistently and is stochastically dominant across *nearly* all scenarios evaluated. The VoF estimated for each economic metric is reported in Table 9.

6.4.3 Lifetime Emissions

Even when formulated for purely cost minimisation objectives, the proposed approach is found to yield significant improvements in system design sustainability, as captured by Figure 6.11. Reductions in estimated lifetime emissions of 54.6%, 67.5%, 52.9% and 56.6% are obtained compared to BD1 and BD2, BD3 and BD4, respectively. Interestingly, the static designs evaluating EH (BD2 and BD4) yield higher net emissions than the others, driven by the UB grid carbon footprint, in line with previous work [247]. This suggests a sustainable ger energy system solution must not rely on the national infrastructure available, but rather on the development of a more distributed and adaptable network, particularly to integrate the multi-energy boundaries encountered. The respective contributions of coal and EH to meeting thermal energy demand for FD over time, shown in Figure 6.11, suggest that most annual carbon footprint reductions can be achieved around year 18. Nonetheless, some emissions are maintained for the entire project duration, as supplying heat with coal or electricity from the UB grid in certain timesteps can reduce costs. The VoF in the context of sustainability is, in essence, the expected avoided 30-year emissions through the FD as compared to the any of rigid alternatives investigated, estimated at 5,532 tonnes CO_2 eq. for this system.

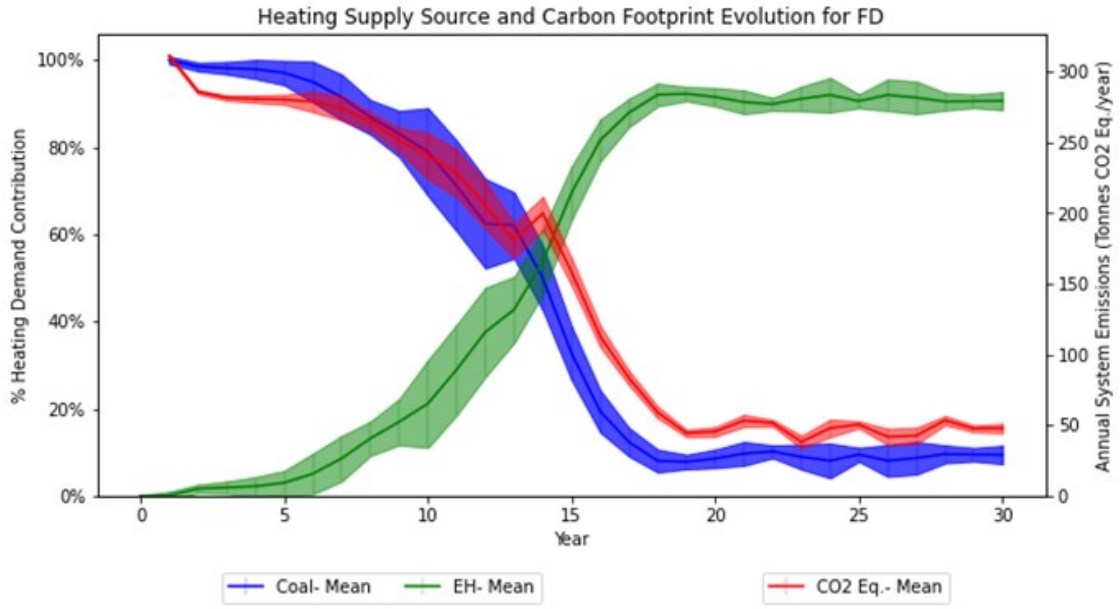


Figure 6.11: Heating supply source (left-axis) vs carbon footprint evolution (right-axis) for FD system over 2000 simulated scenarios showing the 95% confidence interval.

6.4.4 Energy Security and Resilience

Outside of improved economic performance and sustainability, FD also achieves substantially enhanced energy security, with reductions in expected TUL of 76.1% compared to BD1 and BD2, and 74.4% and 73.5% from the TUL values obtained in BD3 and BD4. This is driven by the DRL based flexibility strategy, which can closely match capacity and demand over time for the required provision of energy services. The increased generation capacity for FD also reduces reliance by gers on the UB grid, which can be highly unstable due to frequent load shedding. The PP interconnection allows for the smoothing out of load/generation imbalances as compared to the independent nanogrid units, helping increase system resilience for all stakeholders involved, while also reducing BESS requirements. The VoF for energy security is estimated at 22,804 kWh on average for the duration of the project. Intuitively, this implies that for a simulated 30-year horizon, 22,804 kWh of unmet load could be expected to be avoided by using the FD as opposed any of the alternatives considered. Noting that while some unmet load is still recorded for FD, this results from the allowable capacity shortage presented in 6.2 and the intrinsic stochasticity of the environment investigated, as it would require substantial oversizing of components to consistently meet all loads.

6.4.5 Local Context Constraints

Average energy system expenses (excluding unmet load charges) are around 83 USD $month^{-1}$ ger^{-1} for FD, and substantially lower for BD1 and BD2 at around 52 and 46 USD $month^{-1}$ ger^{-1}, although with significant variation across uncertainty scenarios. The number of timesteps where the budgetary constraints are violated, however, are significantly lower on average for FD. On one hand, this is driven by distributing required investments over project life as compared to significant upfront expenses (and associated replacement costs) recorded for baseline designs. As coal expenses are

reduced over time, the budget for expansion decisions for FD increases, which presents nice synergies with the projected load growth, allowing for the deployment of generating technologies only *if and when* needed. Moreover, the PP enabled FD reduces reliance on BESS, which is both the costliest and heaviest technology considered, helping to limit total system mass.

Accordingly, energy system weight for FD is found to be on average around 15% and 19% lower than BD1 and BD2 for the first half of the project, respectively. The much higher final generating capacities installed for FD and extra mass associated with cabling, however, lead to a net increase in end of project system weight of roughly 28% and 22% as compared to BD1 and BD2. BD3/4 usually result in slightly higher weights than BD1/2, but still within the acceptable bounds. Notably, energy system evolution and expansion possibilities over time are a result of reduced coal mass requirements via EH substitution, similar as what is observed for budget. Given the very large amounts of coal consumed by a typical ger, and the availability of regional lightweight systems, mobility constraint violation is thus very unlikely in FD. The few recorded instances can be attributed to energy generation and demand uncertainty as well as potential load shedding, which can result in additional coal mass requirements from that originally estimated.

6.4.6 Value of Plug and Play (VoPP) Operation

In this section, the VoPP is decomposed from the VoF reported in Table 9 to help assess its economic feasibility, according to (3.15). The results shown in Table 10 suggest it is a worthwhile investment in all cases. These values may be intuitively thought of as the difference in total cost between the same energy system configurations if integrating PP operation or not. FD benefits the most from PP, followed by BD2/4, as the explicit consideration of EH increases system load substantially, and thus motivation for interconnection compared to BD1/3. Nonetheless, for the targeted low-income communities, VoPP for BD1/3 is still significant and can likely be further enhanced. These findings highlight the crucial role of PP in enabling economically feasible mobile energy system solutions, with cost reductions driven by a combination of reduced storage requirements, unmet load penalty and electricity purchased from grid. At the same time, they suggest that a systems approach and flexibility strategy under uncertainty are essential to unlock its full potential, with significantly greater VoPP recorded across all objectives evaluated for FD.

Table 10: Estimated VoPP and externalities on energy security and sustainability for different system design alternatives

Design(s)	**VoPP (USD)**	**% Δ TUL from PP**	**% Δ CO_2 from PP**

FD/FDN	**45,313**	-15.1	-7.3
BD1/3	**11,134**	-6.3	-3.2
BD2/4	**27,312**	-9.4	-5.6

It is important to note, however, that this estimated value is highly dependent on the assumptions used, particularly in terms of herder migration modelling. Significantly greater spatial dispersions than the ones investigated could lead to unsustainable system losses from low voltage distribution, the costs of which may outweigh the benefits generated by PP. To further investigate the effect of this, average distribution losses are plotted against the range of potential cluster radiuses encountered in Figure 6.12. The "break-even" *starting* cluster radius (where VoPP = 0) is estimated at roughly **12.1 km** for the 18-ger system FD through the economic value of unmet load c_{ul}. This implies that in regions with greater spatial dispersions among herders, higher distribution voltages should be considered to maintain an acceptable network efficiency. Alternatively, government policy programs or decision support tools could be developed to incentivize nomadic households to settle closer to each other, helping to capture the full benefits enabled by PP.

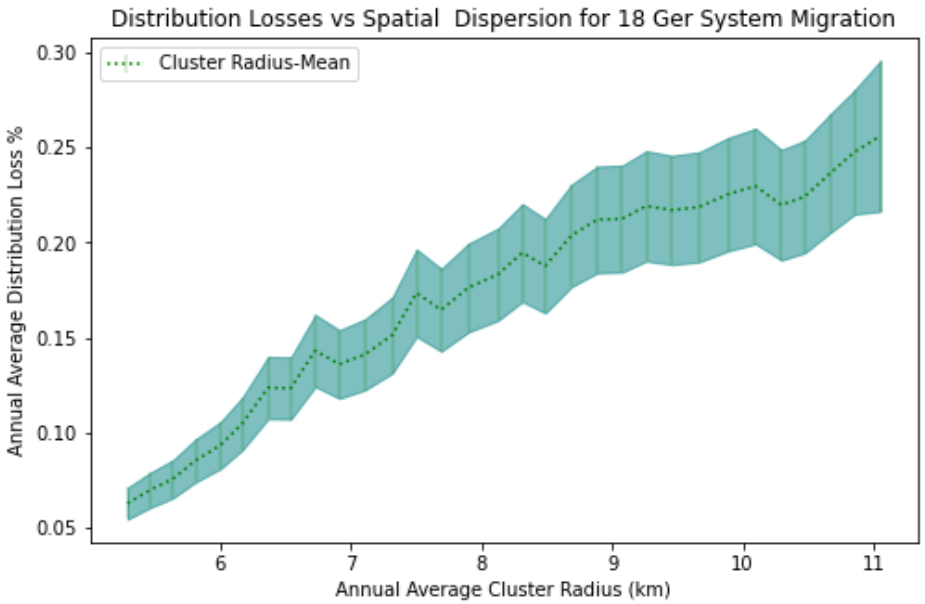


Figure 6.12: Annual distribution loss % as a function of herder PP microgrid spatial dispersion.

6.4.7 Policy Case Sensitivity Analysis

Ambitious governmental policy objectives have the potential to materialize in Mongolia coming years, which may impact the optimality of the energy system designs presented in 6.4. These are primarily based on CAPEX subsidies for RES[202] and EH[247], or participation in carbon credit markets[278,279] through the clean development mechanism (CDM)[280]. In light of the above, 5 additional policy cases (PCs) outside of the baseline are developed in order to investigate performance sensitivity from changing market support instruments. They are formulated to represent different

policy makers-objectives, potential subsidy mechanisms, as well as degree of ambition in meeting sustainability targets. Their focus and modelling parameters are summarized by Table 11.

Table 11: Summary of policy cases investigated for sensitivity analysis of design alternatives considered.

Case	Description	Summary
1	No subsidy for EH, RES, carbon credit	**Baseline (PC1)**
2	WB RES subsidy @ 50%, NO EH financing or carbon credit	**High RES Focus (PC2)**
3	WB EH financing, No RES subsidy or carbon credit	**Clean Heating Focus (PC3)**
4	Carbon Credit @ 6.3 USD 2020 (China ETS) through CDM, no EH or RES subsidy	**China Carbon pricing Focus (PC4)**
5	Carbon Credit @ 35 USD 2020(EU ETS) through CDM, no EH or RES subsidy	**Global Carbon pricing Focus (PC5)**
6	WB EH financing, RES subsidy @ 50%, Carbon Price @ EU ETS through CDM	**Highest Sustainability Focus (PC6)**

The ENPC of the mobile micro-grid is thus recalculated for each of these policy cases (PCs), comparing BD3 and BD4 to the FD solution, given they significantly outperformed BD1 and BD2 under the baseline assumptions (PC1). The results of the sensitivity analysis are captured by Figure 6.13, noting that all designs are optimised based on PC1 only. The graphics shows that the flexible DRL system (FD) achieves the greatest cost reduction across PCs, suggesting its flexibility strategy can generalize to changing support instruments. Additional emissions reductions are also incentivized compared to the baseline case. More specifically, ENPC is reduced by around 10%, 8%, and 17% for BD3, BD4 and FD, respectively. In the clean heating focus, on the other hand, ENPC is reduced by 7% and 15% for BD4 and FD, with no changes in performance for BD3 as EH deployment is not considered. In lack of a more detailed framework at the time of writing, carbon credits received in PC4, PC5 and PC6 are calculated through the difference in emissions from BD3. This leads to reduction of ENPC by 8%, 17%, and 35% for FD, respectively. In line with the analysis presented in 6.4.4, the inflexible baseline including EH (BD4) is actually estimated to generate more emissions than BD3 as it purchases carbon-intensive electricity for the grid to power the heating supply. As such, its energy system cost is not impacted by PC4 or PC5, and reductions in PC6 are driven solely

through the RES investment subsidy. In summary, the results from this sensitivity analysis are promising for the proposed approach, particularly considering that the DRL agent policy could be updated over time as different market support implements are actually implemented.

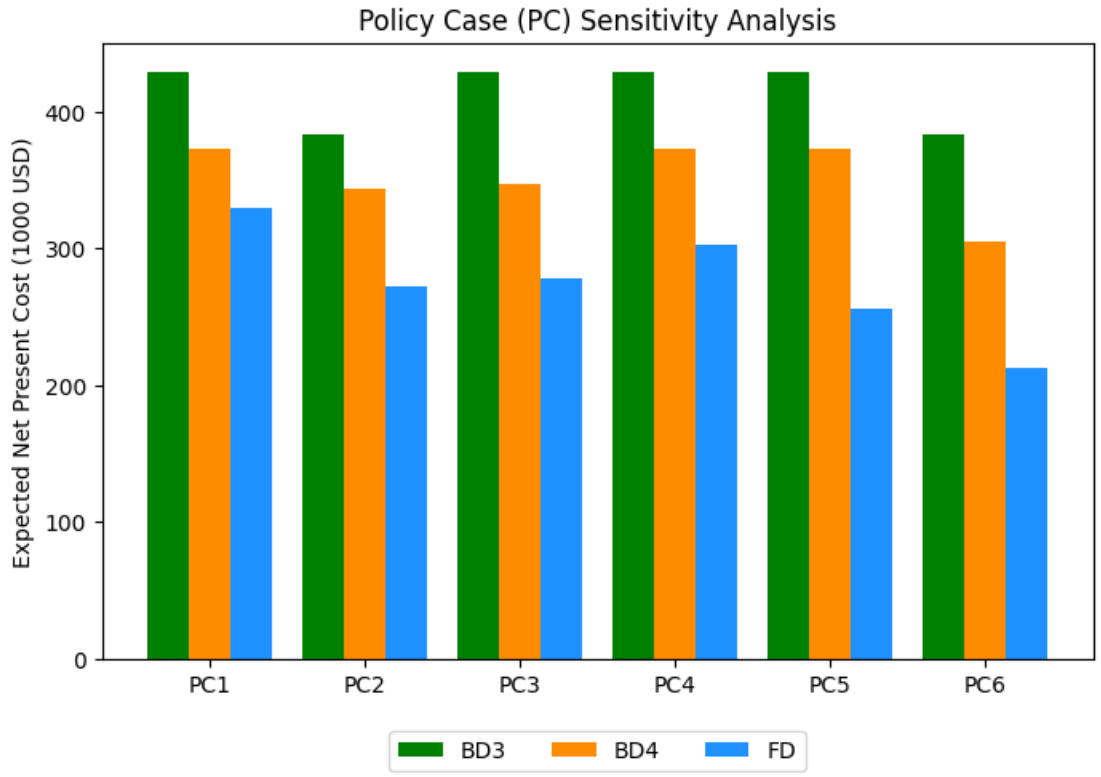


Figure 6.13: Policy sensitivity analysis of primary design alternatives considered in this chapter based on assumptions presented in Table 11.

6.5 Interpreting and Implementing Complex DRL policies under Uncertainty

6.5.1 Overview of Flexible Energy System Evolution Strategy

The FD produced via the proposed DRL based approach suggests the development of a highly dynamic decision-making policy, with capacity evolution based on uncertainty realisations in each scenario. The modular nature of the adaptability strategy is graphically captured in Figure 6.14 below, showing installed nominal capacities over time for different technological components. The plot suggests that in early years the DRL policy presents limited variability, and largely mirrors the findings from the inflexible baselines, with deployment of PV and BESS. Over time, however, it grows increasingly dynamic as more RES capacity is installed within the energy system. This is primarily driven by expansion of wind technologies after about year 7, which then also allows for more EH capacity to be sustainably and reliably operated. Furthermore, in line with near term projections likely to be more accurate, the policy grows more variable over the project life, adapting to the magnitude of uncertainty realizations observed.

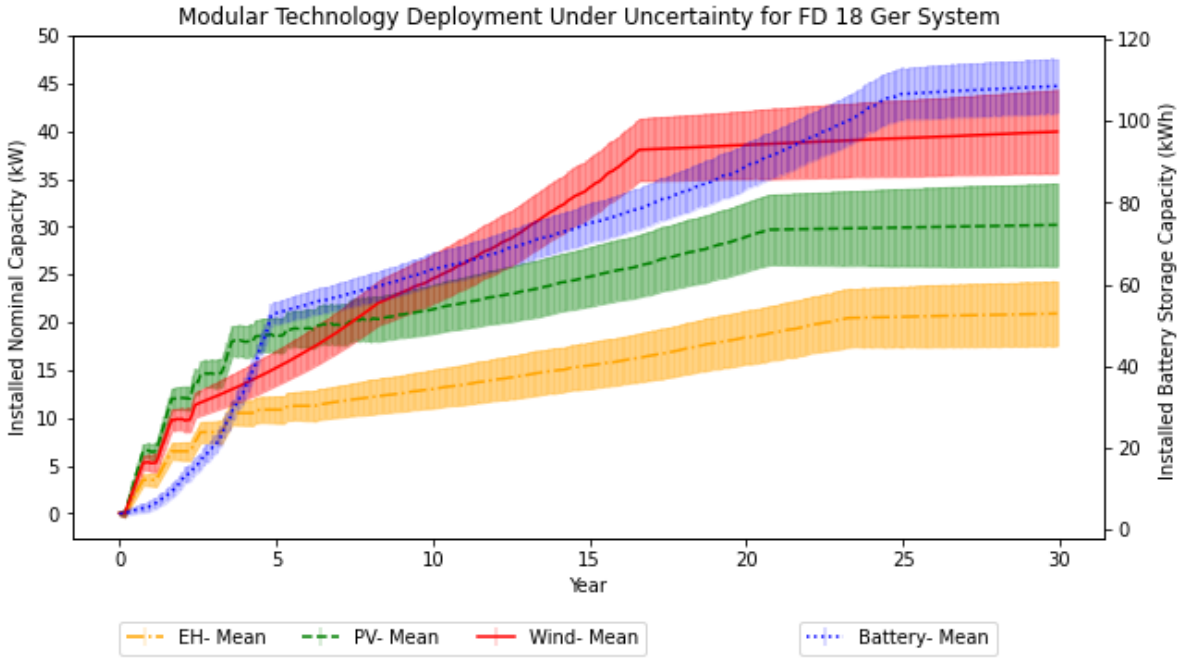


Figure 6.14: Cumulative System Capacity Expansion for the trained ACER agent averaged over 2000 episodes. Shaded areas show 95% confidence interval across scenarios.

6.5.2 Decision Rule Concept Generation and Analysis

Generally, computational insights for decision-making can be generated using the same approaches as in 4.5 and 5.5. Energy system state over time may be used as an input to the DNN parametrized policy, returning either an action probability distribution or a recommended decision as the output. The complex flexibility strategy obtained, moreover, can be further decomposed for improved interpretability by evaluating actions selected under different conditions (akin to *concept generation* for novel flexibility decision rules). Several metrics can be developed for this purpose, through a mix of intuition, domain expertise, statistical analysis, and trial and error. They may then be used to inform options for exercising engineering system flexibility at various stages of the project, based on uncertainty realizations of key variables captured in the state vector.

Figure 6.14, for instance, could be broken down into the expansion timing distribution for each technology, as shown in Figure 6.15 – providing a more detailed understanding of the internal decision-making process. The visualization suggests the development of a dynamic and modular policy, adapting energy system configuration over project life differently for each major component. The profile for wind generally shows a gradually increasing probability, peaking around year 8, and then approximately annealing to zero by year 27. Changes to PV capacity, on the other hand, are more prioritized in early or late years as compared to wind, and are most likely around the midpoint of the project. The distribution of battery capacity adjustments seems fairly well balanced for years 5 to 15, with a gradual reduction in frequency after this. Most varied among these is the temporal profile for EH decisions, which presents a first peak at year 5, and a second one around year 15. This is likely

because the ACER agent waits for RE capacity to rise sufficiently to make an investment in EH worthwhile (i.e. to make sure there enough electricity to operate it).

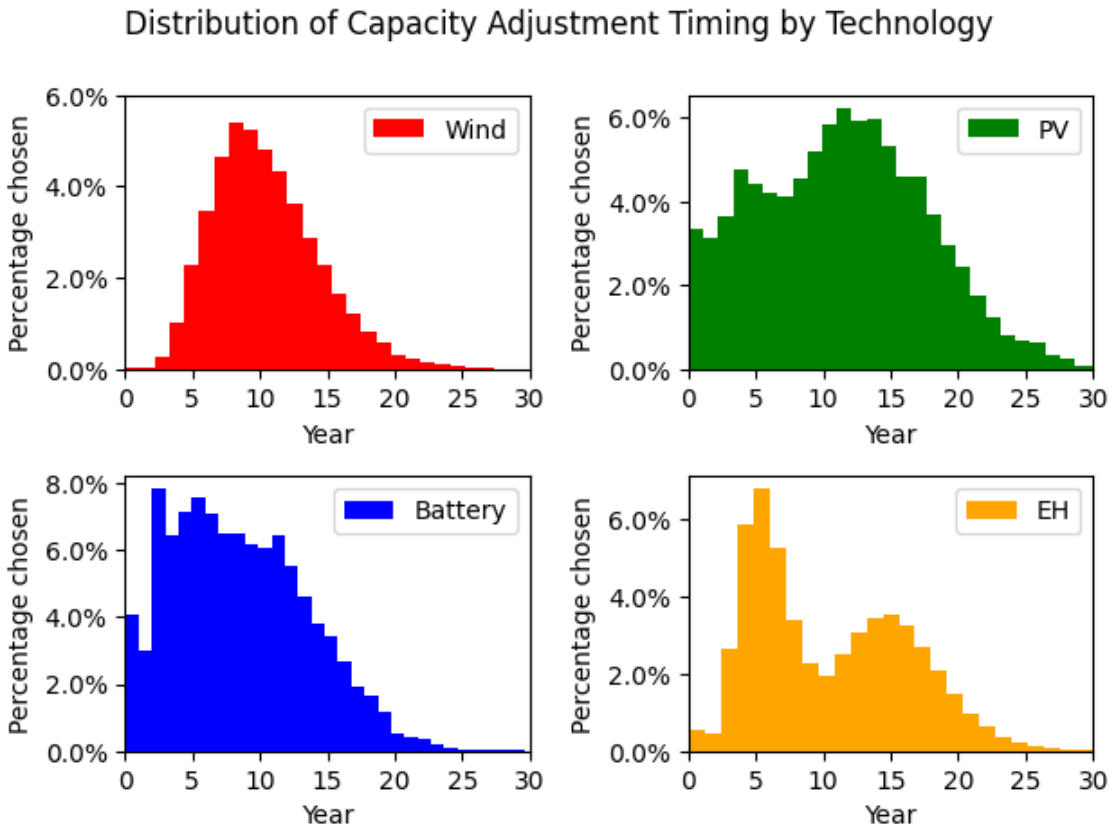


Figure 6.15: Distribution of capacity adjustment timings per technology selected by ACER agent over 2000 simulated scenarios (regardless of magnitude selected).

Another option is plotting the distribution capacity adjustment decisions for each technology by their magnitude, regardless of timing. This presents some parallels to the analysis performed for the proof-of-concept case study in 4.5.2, although it is be remarked that there are no predefined decision rules to compare solutions to here. As captured by Figure 6.16, the profile for wind, PV, and battery capacity adjustments are overall similar. They are most likely to expand by a single module, which can be partly attributed to cash flow limitations of the targeted low-income communities, followed by the max allowable expansion. Abandonment is seldom selected across all system components evaluated, given the overall projected trends in increased energy demand over project life. Interestingly, the distribution for EH differs from the others considered in Figure 6.16, as the max allowable expansion is the most commonly selected action. This can possibly be driven by the ACER agent seeking to reduce the very large coal expenditure(and mass) required for heating supply in the winter months.

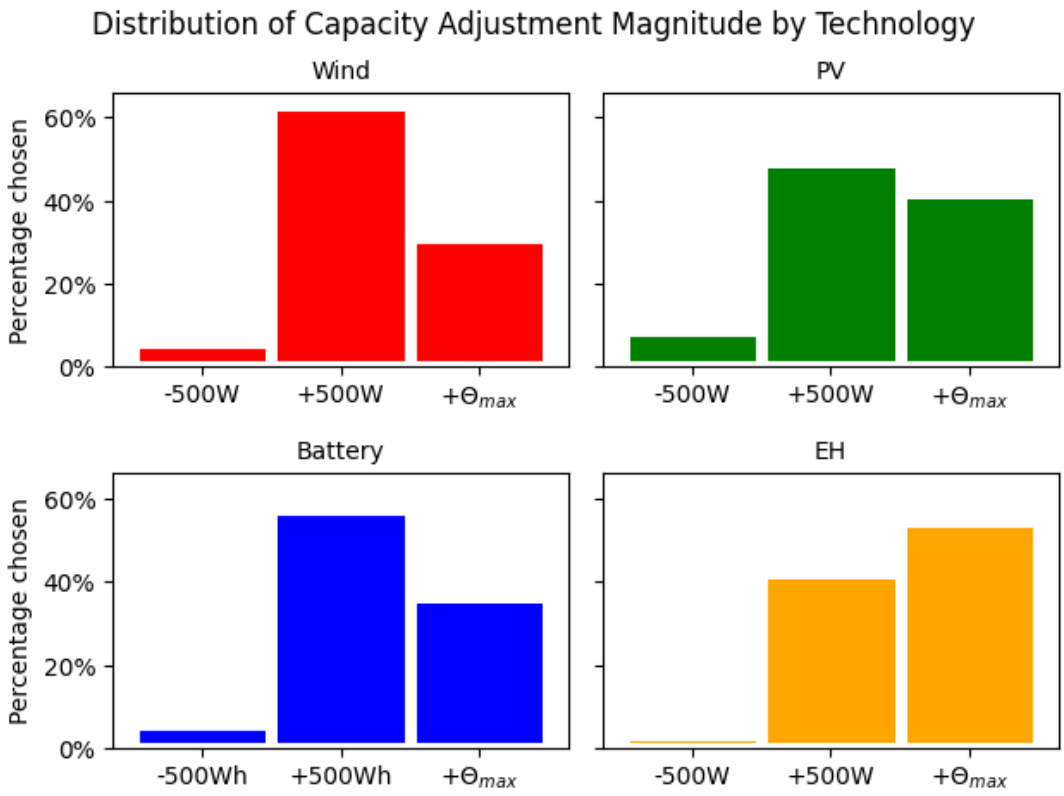


Figure 6.16: Distribution of capacity adjustment magnitude per technology selected by ACER agent over 2000 simulated scenarios (regardless of timing). All plots share the same y-axis depicted on the left hand side. For all plots shown, the left bin represents the frequency of abandonment decisions, the middle bin modular capacity expansion, and the right bin expansion by the max capacity allowed by budget and mobility constraints. As the latter is rounded to the nearest number of modules, it should be noted that this may represent different magnitudes across environment episodes evaluated.

The ratio of installed renewables to EH nominal capacity provides a somewhat less intuitive example metric, as captured by Figure 6.17. Here, a pattern can be observed where this ratio increases at first, then starts to decrease around year 12 and stabilizes at **around 4.2 to 5.1** in the final periods. This could provide a very important guideline for determining energy system evolution, particularly given the heating supply decarbonisation objectives for nomadic ger communities. It may be used, for instance, in combination with the profiles from Figure 6.14, Figure 6.15, or Figure 6.16 to improve understanding and acceptability of triggering signals implemented for capacity adjustments.

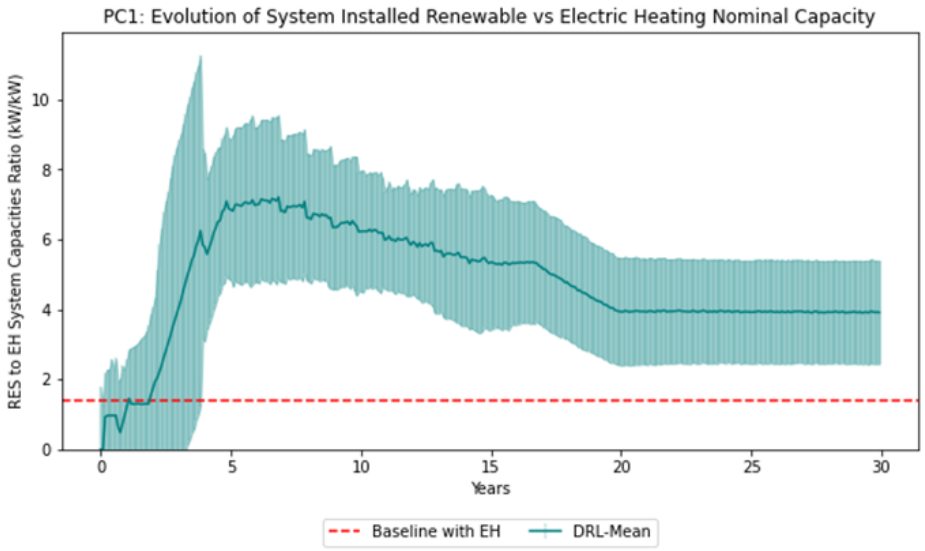


Figure 6.17: RES to EH System sizing ratio for 2000 simulated episodes showing the 95% confidence interval. Baseline with EH refers to BD4 from Table 8. PC1 denotes baseline government policy case (no subsidies or carbon credits).

The behaviour of the DRL agent under different shortage levels also provides some interesting intuition to system operators, as shown in Figure 6.18. It indicates that it is worthwhile to prioritise RES expansion (and wind in particular) under high shortage levels, although important to achieve a balanced generation mix over time. The results also imply that while the majority of EH expansion should happen when there is no unmet load, some capacity should be deployed even in early high shortage years. Most likely, this is because there could still be periods of high EH utilization rate within those years, which can lead to significant cost savings. Energy storage, on the other hand, seems to be prioritized in years with low unmet load levels. This suggests the DRL agent estimates BESS to be particularly helpful at smoothing out those smaller power imbalances, created by intermittent RES generation or stochasticity introduced into the load profiles.

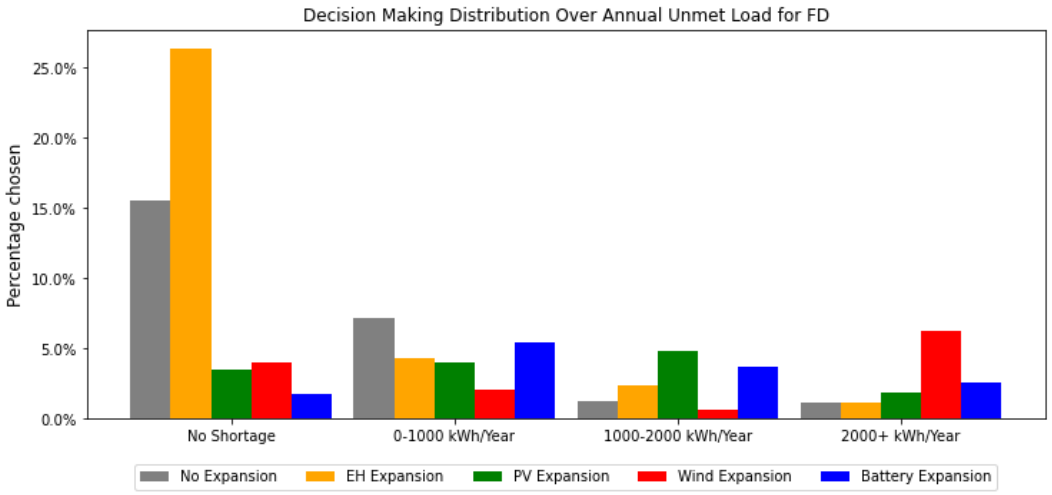


Figure 6.18: Distribution of capacity expansion decisions vs shortage levels over 2000 episodes for the trained ACER agent.

6.5.3 EnerGer: Design Decision-support Tool

The retroactive implicit decision-rule distribution presented in the former section already provides some interpretable guidelines which could be used by herder communities to determine their energy system expansion strategy. To capture the full value and dynamicity of the proposed approach, however, a decision support tool could be implemented. This would be targeted at improving data visualisation, understanding of the policy optimisation performed, and ultimately enhancing acceptability. A prototype mobile app named "EnerGer" was developed for this purpose by Alex Gibson, captured by Figure 6.19, which was based on extensive collaborations as part of his Master's thesis in Design Engineering[281] .This can help community members track energy system state, notify users of infrastructure replacement requirements, explore locations of other nearby PP during migration, and dynamically recommend actions over time. Given the communication infrastructure is already present, it could be feasibly implemented at a low cost. This approach presents many advantages to better tailor energy system solutions to each customer. Users with different budgetary or mobility requirements from those assumed in this work, for instance, could enter that data directly in the app, which could then call for DRL agent training with the updated parameters. Similarly, actual realized demand for heating and electricity, or recorded energy generation by source, could be manually input over time to produce the most relevant recommendations for exercising flexibility. Furthermore, the development of any market support instruments such as those from Table 11 could be automatically integrated over project life, and system operators promptly informed of their impact.

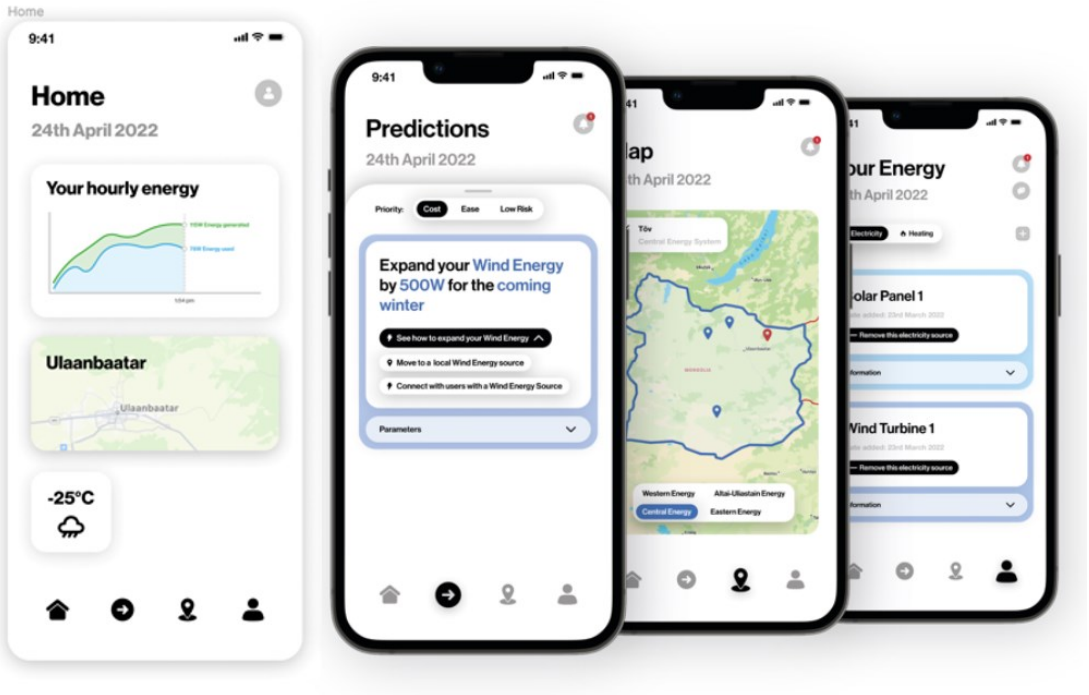


Figure 6.19: Visualization of interface for EnerGer prototype [281]. Going from left to right, it shows the homepage, DRL model predictions/recommendations, a geospatial feature to identify nearby communities during migration, and a summary of energy usage broken down by heating or electricity.

6.6 Discussion

6.6.1 Changing Economies of Scale

Flexibility in Design is generally motivated by tensions between EoS and modular adaptable deployment, where each can be advantageous depending on the setting. The results presented in 6.4, however, suggest that the flexible solution is stochastically dominant and outperforms the baseline designs in nearly all scenarios evaluated. This may be due to the relatively low magnitude of EoS assumed in the cost calculations. Coal expenses, even if significantly reduced, remain the strongest driver of ENPC for nomadic energy systems in Mongolia. While evolving costs for other technologies may influence this finding, an important advantage of the proposed DRL approach is precisely the simplicity to retrain the agent with updated assumptions as they are verified over time. As such, if governmental programs are developed which influence the optimal solution (i.e. subsidy schemes, carbon credits, etc.), they could be quickly integrated into the model to give up to date recommendations. Finally, it is worth noting that removing the budgetary constraints formulated could lead to the discovery of better performing solutions which take advantage of both EoS and flexibility. While this constraint was relaxed in the zero-stage capital expense decision, there are likely to be some practical, logistical and/or economic benefits to ramping up and concentrating investment in certain time periods.

6.6.2 Decarbonisation Pathways

The trade-offs between costs and emissions for inflexible ger energy systems not considering PP were further evaluated through the Non-dominated Sorting Genetic Algorithm (NSGA-II) to the optimisation of BD1 and BD2 [282]. The pareto frontier obtained suggests that high EH capacities should be deployed even by inflexible nanogrid units seeking to minimise ENPC, as they help reduce coal expenses over time - instead purchasing heavily discounted (particularly at night) electricity from the UB grid. The associated carbon footprint, however, means those solutions are estimated to lead to a net increase in expected system emissions over the project horizon. The results presented in the former sections, in fact, suggest that an economically feasible decarbonisation of ger community energy systems is possible even when accounting for the associated uncertainties, due to several important synergies. On one hand, ger communities present substantially higher demand for electricity in the summer months compared to the winter ones, arising from increased refrigerator energy requirements. This allows RES generation to be utilized almost exclusively for EH in the winter months. The relatively high capital expense of RES and EH for low-income communities, as well cultural value associated with the coal stove, nonetheless, remain important limiting factors to the rate of future feasible decarbonisation in Mongolia.

6.6.3 Plug and Play Mobile Microgrid Technical Feasibility

While this chapter was focused on the planning rather than operation side, the new control algorithm developments remain largely untested on the field. This is an important issue to be addressed,

particularly in terms of overall resulting network efficiency and feasibility at different levels of spatial dispersion. In this study, a ring like interconnection was fixed, however this is likely to vary significantly during actual herder migration, which may strongly impact the performance of a design and flexibility strategy. Additionally, even though only active power was considered here, there are substantial safety and reliability issues associated with low-voltage AC distribution over long distances. Nonetheless, the results from this work imply that there is a significant economic value attached to PP operation, thus motivating the need for more in-depth feasibility studies.

6.6.4 Limitations and Opportunities for Future Work

While this work presents an important starting point to evaluate the planning of future nomadic and mobile energy systems, there are some other limitations which should be noted. Data on load profiles and projections were difficult to find, and several assumptions had to be made to develop the model from 6.2.1. A higher resolution heat load profile model, perhaps based on outside temperatures in each hour per day, could allow for a more detailed evaluation of different climate change scenarios. The mobility aspect within the energy system simulation is completely novel and the modelling approach implemented hard to validate. Continuing interactions with local energy sector stakeholders and herder communities will be crucial in this regard. Especially valuable insights may be on the flexibility of migration patterns assumed, and willingness to deviate from them to engage in PP operation in different regions. Additionally, the results obtained are highly dependent on realisations of different cost factors over the project life. These were mostly assumed to be invariant in time for simplicity, and for lack of better information, however, are highly likely to differ over the next 30-years. In order to address data limitations at the time of writing, future work ought to investigate reproducibility of solutions when using different stochastic model parameters. Considering the possibility of new bulk procurement programs, it may be interesting to policy makers to perform a more in-depth sensitivity analysis on the impact of EoS in Mongolia. This can suggest what the break-even values may be, or where the rigid designs outperform the flexible ones.

The action space for the DRL agent could be significantly expanded from this implementation to reflect the relevant decision-making possibilities. Multiple technologies could be considered for expansion at once, for instance. Moreover, connection to the UB or a nearby grid could be explicitly modelled, and respective recommendations produced. Several other potential actions were not evaluated here, particularly higher efficiency stoves or improved ger thermal insulation, which could be highly valuable for decarbonisation objectives. Inverter capacity was also not explicitly analysed, thus likely leading to sub-optimal operation at different points of the project. Closer integration of the control and planning aspects of this problem should also be investigated in future work. While the computational expense required may make the problem intractable, it would be valuable to actually simulate PP operation at a finer time resolution over the entire project horizon. Alternatively, a separate DRL agent could be used to develop more advanced operational strategies than the LFS

implemented here. Flexible demand response and prioritised load shedding, for instance, are likely to help reduce ENPC incurred by mobile energy systems.

From an algorithmic point of view, there are several more interesting avenues for future research. More specifically, a multi-agent formulation could be developed to better evaluate the costs and benefits to individual herder households, and the resulting optimal flexibility strategies. Recent advancements in distributional DRL allow the formulation of a policy based on the entire distribution of rewards rather than purely expected value [142]. Recommended strategies could thus be tailored to risk-seeking or risk averse decision makers, improving acceptability, and understanding. Finally, exploring alternative DNN architectures which are better tailored to the tackled problem could lead to further improvements. CNNs could be used to map relationship among features which may not be intuitive at first, for instance, while RNNs could help capture the seasonal nature of decision making required in order to better adapt agent policy between winter and summer months.

6.7 Chapter Summary and Conclusions

In this chapter, the proposed DRL-based approach was implemented for the design and planning of a mobile energy supply system, based on a case study on Mongolian herder communities. Motivated by the potential for increased climate induced migration and increasingly decentralized generation, a novel modelling framework to integrate mobility considerations was developed. Contrary to popular energy system planning methods, the data-driven approach used allowed for the tractable integration of several sources of uncertainty as well as mobility within the evaluation model. Both heat and electricity were considered, leading to a more holistic evaluation of different solution alternatives than previously available. The design and planning strategy for a highly flexible and dynamic energy system was thus optimised through an actor-critic algorithm, with a reasonable computational expense. Benchmark inflexible designs with economies of scale were compared across key economic, sustainability and resilience indicators such as Cost, Equivalent Emissions and Total Unmet Load. The results suggest that the proposed approach can lead to economically feasible mobile energy systems, even in the case of budgetary and mobility constraints. Additionally, the value generated by "Plug and Play" operation was also estimated using a variation of real-options theory, suggesting it is a worthwhile investment in all cases. As such, this chapter provides an important contribution to enabling energy access for nomadic communities worldwide, and general energy systems planning.

The analysis presented here addresses several key limitations from Chapters 4 and 5 in order to best complement the field of Flexibility in Design. As a starting point, the proposed approach was shown to be effective even on an extremely highly dimensional case study subject to several sources of uncertainty. The optimization of the flexibility strategy performed, in fact, would be computationally prohibitive using standard methods if accounting for the full uncertainty space. Given the lack of comparison with a flexibility decision rule implementation, novel approaches were explored to

improve the interpretability and acceptability of the methodological contributions of this thesis in practice. These ranged from visualizing the overall energy system cumulative evolution strategy, to decomposing it into capacity adjustment distributions by timing or magnitude for each technology. Novel metrics and potential decision rule variables were developed as a result, including the ratio of RES to EH and action selection based on yearly shortage levels. Furthermore, an initial prototype for a design decision-support system based on DRL flexibility analysis was presented, aiming to improve the practicality of using recommended strategies in real-world system operation. Finally, a policy case sensitivity analysis was performed, suggesting a strong generalization ability for the DRL agent. Nonetheless, it should be noted that all market support instruments were beneficial to reducing costs, leading to *biased* out of sample testing, even if DNN policies were not re-optimized based on the updated environment assumptions. In light of the above, the methodological maturation achieved in this chapter provides a blueprint for how the proposed approach can be implemented on practically any engineering system which fits into the DRL flexibility analysis framework. It should still be noted, however, that a more complex system such as the one tackled here inherently carries more risks. Challenges arise from the many variables and interactions among system components, increasing the need for validation, coordination among technical disciplines, and rigorous methodological implementation. Accordingly, before deploying these solutions in the real world, further validation and testing is recommended.

Chapter 7

7 A Computational Framework for Data-driven Formulation of Policy-making Tools under Uncertainty

This chapter builds upon the methodological maturation presented in Chapter 6, by extending the contributions of the approach developed in this thesis to potential government policy-making objectives. The sensitivity analysis of the DRL based design to market support instruments detailed in 6.4.7 motivated this section. In essence, it inspired the key question of:

- RQ7: What other valuable application areas are possible with the proposed DRL approach outside of direct flexibility evaluations?

In order to address this, a novel computational framework for data-driven optimization of policy-making tools under uncertainty is proposed. Different policy mechanisms are evaluated to accelerate and enhance the achievement of SDG7 for global nomadic communities through PP enabled mobile energy systems. Mongolian ger communities are used as an example case study as this issue is highly relevant to governmental objectives there, although the approach and results can be easily translated to other locations through the adjustment of the appropriate parameters. Many other migratory communities exist across several continents, for instance in Nigeria, Africa[283], Argentina, South America[284], and Kazakhstan, Asia[285]. Their energy usage patterns are likely to exhibit many similarities, capturing the urgent need for decarbonization of heating and cooking supply sources, while pursuing energy access and developmental objectives.

The framework presented here was also developed with generalizability in mind, and thus is compatible with the formulation of policy tools such as subsidies or tariffs for any energy system to which they are applicable, not exclusively mobile ones. Contrary to the previous case studies, however, the primary perspective considered is that of government or institutions looking to design effective policy tools for their objectives. Accordingly, a modified testing procedure was implemented than in Chapters 3-6, aiming to decompose the understanding of the quantitative impact for each policy-tool considered. The most significant novelty from previous works is, in essence, the ability to evaluate the impact of a much larger policy-tool uncertainty. Furthermore, this analysis extends past cost, sustainability, and resilience measures to also consider energy access metrics.

This chapter is partly based on a manuscript submitted to Nature Energy which is currently under review, with Pudong Ge and I as co-first authors[286]. Pudong was responsible for the section detailing PP operation, while I developed the rest of the material presented in this Chapter. The full open source code to reproduce all results presented in this chapter can be accessed here[217].

7.1 Case Study Overview

Previous achievements towards the United Nations Sustainable Development Goal (UN-SDGs) 7 – "*Ensuring access to affordable, reliable, sustainable and modern energy for all"* -have significantly slowed down in recent years as remaining targets are increasingly harder to reach, with 733 million people still lacking electricity access[287]. They are located primarily in developing or lower-income countries, where the costs of centralized grid extensions are prohibitive, especially for very remote areas or communities following a migratory lifestyle[288]. Surprisingly, however, the UN-SDGs do not make a direct connection between climate change, migration, and energy access – even considering estimates that up to 1 billion people may experience climate change-driven migration by 2050[289]. Currently, at least 40 million rural nomads exist globally, whose livelihood is ecologically adjusted to pursue agricultural activities in areas with limited natural resources[290]. Typically following seasonal movement patterns using livestock and some form of mobile dwelling to search for better pastures, rural nomads are likely to be severely impacted by the ongoing changes to the climate (i.e. precipitation, weather patterns) in coming years[289]. Additionally, recent studies suggest that, global reductions of areas where sedentary agriculture can be reliably practiced will lead to around 143 million more people practicing rural nomadism by 2050[291,292]. Due to many of the same factors, recent years have experienced a rapid rise in climate driven relocations to informal settlements on the outskirts of major cities[293]. Primarily as a result of grid infrastructure limitations, over 200 million urban migrants are completely unable to access electricity or heating services, impacting many other developmental objectives[293,294]. Moreover, the limited disposable income of these communities means they still tend to rely on the inefficient combustion of biomass for cooking and heating, with significant negative pollution externalities[56].

Consequently, there has been growing interest in decentralized solutions, with mini-grids and micro-grids systems expected to be the main driver of future rural electrification, currently providing energy services to around 140 million people[56,57,295]. Their techno-economic feasibility has been abundantly explored in the literature[296,297], with recent work suggesting that high levels of reliability and decarbonization can be achieved through renewable energy and battery energy storage (BESS) systems[58,60,229], even without a dispatchable generator. Recent developments in Plug and Play (PP) control algorithms based on a sparse and hierarchical communication framework[298], nonetheless, can exploit advantages of both centralized and decentralized electrification strategies. The temporary interconnection of nearby households allows the formation of mobile energy supply systems, improving compatibility with a nomadic lifestyle, and increasing the scope of the energy transition.

Moreover, PP operation presents many sustainable electrification benefits for both low-income communities and centralised power grid operators. It allows the smoothing out load/generation mismatch from a household to a community level, and an additional energy supply source with market engagement opportunities. This can also lead to enhancements of main grid reliability through the integration of more distributed generators and storage[225,299,300]. Finally, depending on local climatic conditions, PP may drive the decarbonization of nomadic energy systems by promoting a transition to electric heating or cooking powered by emission-free renewables.

Given these pressing issues, developing energy system modelling, planning and optimization frameworks which are compatible with a migratory lifestyle is essential to achieve SDG7. Common planning approaches typically require substantial simplifications to tractably reach a solution, typically by reducing the associated scenario or technological boundaries[301,302]. The extensive long-term operational uncertainties and budgetary constraints faced by these communities, therefore, necessitate novel data-driven approaches to plan strategies for flexible systems which modularly deploy capacity over time[10,61,117].

The mobile energy supply system considered here alternates between stand-alone/rural community operation in the summer and urban migration with connection to the main grid when available in the winter. Accordingly, three different potential common policy tools are compared and optimized based on SDG7 metrics[303]: (a) investment subsidies; (b) feed-in tariffs; (c) carbon credits. Note that these were refined from the original policy case formulation in 6.4.7. The method scales well computationally and can easily be adapted to assumptions. The aim is to identify which mechanism is the most effective at achieving each energy access metric in this context to quantitatively inform future governmental policy making based on their targeted objectives in specific scenarios.

7.2 Background and Previous Work

This section presents a description of the policy-making case study assessed in this chapter. The modelling approach and background are similar to the ones introduced in 6.2.1 and 6.2.2 respectively, with any differences to original assumptions or methods explicitly stated. The following provides a succinct overview of the governmental policy formulation problem in light of different decision-makers perspectives, thus minimizing any repetition of the material from Chapter 6.

7.2.1 Nomads, Climate Change, and National Energy Policy Objectives

Mongolia presents an excellent case study to evaluate the global issues above, given that 25-40% of the population follows a nomadic lifestyle, moving up to 30 times a year[304]. Further, over the past 50 years, the average temperature in Mongolia has risen by 2.24 °C, much higher than the global average[305]. This has exacerbated a periodic weather phenomenon known in Mongolia as the *dzuz*[306], a harsh cold winter that follows an unusually dry summer. Many nomadic settlements (or *gers*[207]) were originally pursuing a fully rural lifestyle, but were driven to settle down through urbanization[191,197,203]

from changes to climatic conditions impacting their livelihood. Currently there is still a small population following the traditional lifestyle, with most nomads moving around grasslands in the summer and settling down in ger districts[191] during winter months. Consequently, the available infrastructure cannot meet the needs of the overextended informal settlements, furthering development issues associated with nomadic energy access, and promoting excessive fossil fuel consumption[192].

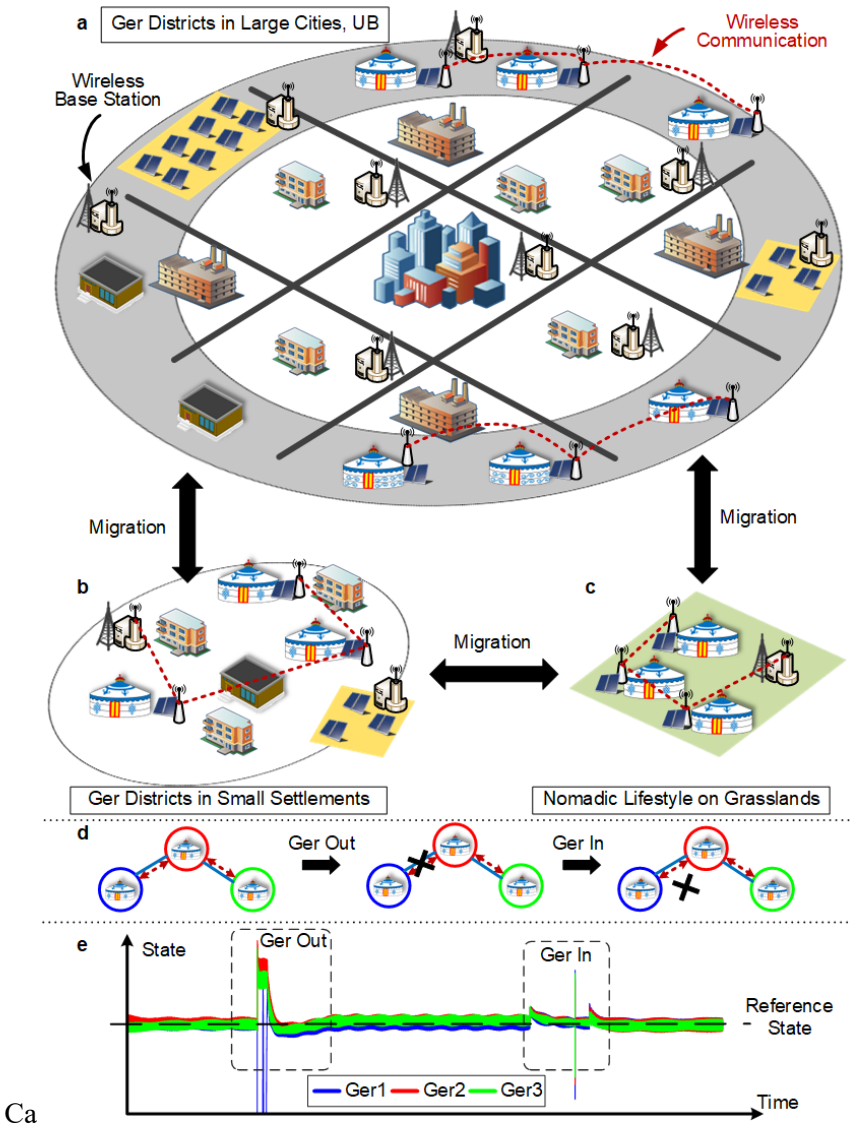
In fact, gers still primarily use indoor coal-burning stoves for meeting thermal energy demand during the long and harsh winter, causing severe carbon emission and air pollution problems. As a result, annual mean particulate matter ($PM_{2.5}$) in Ulaanbaatar (UB) is 6 to 10 times higher than the safe air quality level guidelines, and is one of the strongest drivers of infant deaths[307]. However, the outdated energy network infrastructure only covering a fraction of the potential ger migration range[205,206,252] meets excellent renewable potential, allowing a low-carbon solution. As a result, there have been several institutional and governmental programs that have contributed greatly to progressing on SDG7 for nomadic communities in Mongolia. The internationally sponsored Renewable Energy and Rural Electricity Access Project (REAP) had the largest impact, providing investment recovery or capital expense (CAPEX) subsidies[202].

Similar to REAP, the Government of Mongolia also has implemented ambitious initiatives electrifying the energy use to reduce air pollution in UB, including a night-time electricity policy for selected ger districts to encourage a switch from coal to electricity for space heating[308], and ger district re-development project[207,309]. Combined with an economic downturn since 2015, however, the heavily discounted night-time tariff created a significant budget deficit[190]. Hence, alternative future policy support instruments should be holistically evaluated across energy access metrics while prioritizing decarbonization objectives. Other relevant policy tools currently in place include the Mongolia State Policy on Energy 2015-2030, Renewable Energy Law and Mongolia Renewable Energy Feed-in Tariff -which varies for on vs off-grid installations[200,206]. The National Green Development Policy and the Sustainable Development Vision 2030 further outline Mongolian desired SDG transition[310]. Most related to SDG7, the ambitious plans aim to achieve a 30% renewables share in the energy generation mix by 2030, and provide up to 75% of herders with the electricity, techniques, and technologies capable of supporting businesses[206].

7.2.2 PP Operation as an Enabler of Urbanized Migration and Nomadism

Renewable based electrification in nomadism settlements should enable PP operations, which correspond to different scenarios in Figure 7.1. Hence, the seasonal migration route considering the mobility is depicted in Figure 7.1 (a-c) where the renewable usage scenarios are classified as static and mobile: large-scale centralized generation (e.g., wind farms or solar PV stations) and mobile light-weight units (e.g., Small Home Solar, SHS). Along with herders moving between ger districts

and remote communities, the control modes of power-electronic-enabled[311] mobile energy units can be seamlessly switched between grid-connected and islanded modes, e.g., PP operations that guarantee power qualities (e.g., frequency stability, voltage stability shown in Figure 7.1e). The short-time-scale (seconds or even shorter) stability issues of PP operations can be mitigated[312] through a flexible wireless communication, e.g., cellular network shown in Figure 7.1. Hence, PP operations of mobile energy units create an operational flexibility opportunity of creating an energy source-demand balance for long-term analyses of technological capacity investments to address energy access objectives in the context of migration, and further in recent trends of increasing distributed loads and generation.



Ca

Figure 7.1: Migration pattern of herders and mobile energy access scenarios. **a**, the city urbanization process with increasing informal settlements around the suburban area (ger districts in the Ulaanbaatar, Mongolia as an example): apart from the commercial buildings, industrial factories, and residential households that all consume energy, urbanization leads to more ger districts around the city centre without enough energy supply infrastructure. Renewable is promising in electrifying energy usage of increasing citizens, both centralized PV stations and decentralized household PV systems.

The control of all these energy units can be assisted by base station enabled wireless communication, while only communication links for Gers (in red dotted lines) are shown. **b**, a similar urbanization process but in some small settlements such as small cities or villages: there are some non-ger and ger households in the residential areas, and mobile PV panels assist the energy usage in Gers. **c**, nomadic lifestyle with increasing energy access by mobile renewable PV systems. For three scenarios (shown in **a**, **b**, **c**), the pastoralists can migrate across these three scenarios, while their mobile energy supply facilities such as PV systems and small wind turbine systems have a functionality of plug-and-play operation. **d**, a simple case of plug-and-play operation with the energy unit plugging in-and-out. **e**, the state (such as frequency and voltage variables in electricity systems) evolution of plug-and-play operation shown in subfigure **d**: with some transient, the state tracks the reference. By the plug-and-play operations (**d**, **e**) of energy supply units, the energy usage pattern in global pastoralists with migration patterns (**a**, **b**, **c**) can be mitigated, leading to an efficient, green, decarbonized energy supply and a healthy living environment with reduced air pollution.

7.2.3 Energy Transition Progress

The impact of implemented energy policies on different metrics and population types has not been consistently reported in the literature thus far, increasing the complexity of focusing on rural nomadism or urbanized migration for different metrics. REAP is deemed to have largely achieved access to Tier I electricity (SDG7.1.1), with a reported 95% value as of 2020, even in rural areas[205]. The SHSs installed, while providing an important first step, are likely insufficient to meet the expected electricity demand evolution over the next 30 years (5-7% annually) – particularly if considering a transition to sustainable heating[204]. In fact, herder electricity consumption currently only makes up about 1% of the national total, even though they account for at least one quarter of the population [206]. Furthermore, the required investment through REAP consisted of a third or more of annual incomes for 40% of herder households[313]. Reported as a significant limitation by many herder families, it highlights the importance of integrating budget constraints and modular deployment strategies[10] toward accelerating the energy transition for low-income communities[202].

Renewable total final energy consumption share (SDG7.2) is significantly below target levels at 3.3%, although renewable makes up 9% of the electricity generation mix[205]. Note that this refers to the entire country, as values specific to ger communities are not publicly available. Recent results[261] on a very similar system to the one investigated here and the prevalence of renewable for herders from REAP, however, lead to an estimated value of 31% on SDG 7.2 for the average ger SHS system. On the other hand, driven by the economic potential to become an electricity exporter, renewables capacity per capita (SDG7.b.1) has improved significantly in Mongolia, achieving 84.4 W per person at the time of writing[205]. This value is driven mainly by a few centralized large-scale installations, thus the capacity in herder communities is limited at 34.2 W per person[202].

In this work, given the PP operations developed and its associated rural electrification benefits, indicator SDG 7.1.1 is assessed through the Multi-Tier Framework and access to Tier III electricity[314], referred to as SDG7.1.1*. This was motivated by recent surveys[313], where around 88% of herders reported not being able to consistently meet electricity demand (mostly in winter when there are fewer sunlight hours) – as well as corresponding load evolution projections[190]. Furthermore, 16% of herders

are reported to have no power supply at all. Accordingly, it is estimated that roughly 4.3% of rural nomadic communities in Mongolia may have achieved Tier III electricity access (SDG7.1.1*) as of 2022[314]. Similarly, in light of the extreme climatic conditions experienced in Mongolia, leading to space heating demand being much larger than that for cooking, access to clean thermal energy (SDG 7.1.2*) is defined as reaching level 5 on the Multi-tier Matrix for Access to Space heating[314]. Progress on this front has been much slower due to cultural factors associated with raw-coal-burning stoves, the carbon footprint of the main electricity grids, and institutional limitations on access to the district-heating network in UB. As a result, only around 2% of gers currently have access to clean thermal energy for heating[315]. Please refer to the policy impact study by [316] for a detailed procedure on how SDG7* indicators were computed and their meaning in this application case study.

7.2.4 Policy-Case (PC) Formulation

Based on interactions with local energy sector stakeholders[190,199], governmental decarbonization objectives[200,256] , and institutional reports[202,205,206,317], four new policy cases (PCs) are developed to estimate the uncertainties introduced in this Chapter. They are described in detail by Table 12 below. CAPEX subsidies are analysed separately for renewable and electric heating given the urgency of the air pollution issue in UB and the very large carbon footprint of its distribution grid[247]. Uncertainty is integrated for each PC through GBM models, generating stochastic scenarios according to (3.11)-(3.12). Essentially, this means that the values for renewable energy subsidies, electric heating subsidies, feed-in-tariffs, and carbon credits are all stochastically simulated. In each 30-year evaluation scenario, therefore, different realizations of all uncertain variables from Table 12 can be observed. These are estimated from values at project start (2022) and reasonable estimates for end of horizon (2052), as well as through the above interactions. Additionally, it should be remarked that while flexibility strategies are evaluated monthly, values for PC1 and PC2 are assumed to be updated every 5 years. PC3 and PC4, on the other hand, are simulated yearly to be more realistic to practical policy-making objectives.

Table 12: Summary of Policy Cases (PCs) analysed, and the ranges considered for DRL agent training. Note that these were refined from the ones in 6.4.7 in line with SDG7 objectives.

Policy Case (PC)	**Description and Projections [Low, High]**	μ	σ	dt (years)
Baseline (PC0)[200]	No subsidies	n/a	n/a	n/a
Renewable Subsidy (PC1)[206,318]	[10%, 50%] of CAPEX	0	0.3	5

Electric Heating Subsidy (PC2)[247,261]	[10%, 80%] of CAPEX	0	0.4	5
Renewable Feed-in Tariff (PC3)[190,205,206]	Wind: [0.10, 0.2 USD/kWh] PV: [0.2, 0.4 USD/kWh]	0.1	0.05	1
Carbon Credit from China ETS (PC4)[280,317,319]	[6.7 2022 USD/ tonnes CO_2 Eq. in 2022, 27.3 2022 USD/ tonnes CO_2 Eq. by 2052]	0.0567	0.1	1

7.3 DRL Implementation

7.3.1 DRL Flexibility Analysis Formulation

This section provides a brief overview of the data-driven approach implemented for policy -making objectives in this chapter. The steps required for DRL flexibility analysis are summarised below, although please refer to 6.3 for a more complete description methods undertaken and motivation.

1. **Environment Formulation:**

   a. State Space: According to Chapter 6, the engineering system evolution strategy is primarily dependent on the periodic demand for different energy services, installed generating capacities, and migratory patterns, among other factors. In the formulation presented here, however, the realization of the most relevant policy tools over 30 years is explicitly included as part of the environment observation. The updated state vector is defined through a box observation set according to (7.1), further increasing the dimensionality from that in (6.25). All state variables are normalized based on the maximum and minimum values obtained for them over 10,000 simulated scenarios, independent of the environment. Time limit awareness $T-t$ is included.

$$s^t = \begin{bmatrix} ED^t, HD^t, r^t_{cluster}, {L_{cb}}^t, {\theta_{pv}}^t, {\theta_{wind}}^t, {\theta_{eh}}^t, {\theta_{bess}}^t, ED^t_{ul,es,final}, M^t_{es}, RE_{subsidy}, \\ EH_{subsidy}, RE_{feed\ in\ tariff}, p_{CO_2}, T-t \end{bmatrix} \quad (7.1)$$

   b. Reward Function: Based on the energy access objectives of this work, the focus is on minimisation of cost needed to provide electricity and heating services to the targeted low-income communities. Accordingly, the definition of reward function from (6.26) is not modified here, including the penalty for violating budgetary or mobility constraints. The policy tools described in Table 12, however, alter the shape of the reward function by changing model parameters and assumptions, generating

additional positive cash flows. For instance, investment cost C_{inv} in (6.11) is reduced in PC1 and PC2 through the subsidies considered, lowering their respective technological deployment cost coefficients (K_{PV}, K_{Wind}, K_{EH} in (6.14)). The impact of feed-in-tariffs in PC3 is captured by changing values for variable p_{grid}, which determines C_{grid} in (6.21). Carbon credit revenues in PC4 are instead estimated from the avoided emissions compared to the most sustainable baseline nomadic community energy system (i.e. BD3 from Chapter 6). As in 6.4.1, BD3 is very representative of the typical system available to Mongolian herders. The reward function form variation is then based on evolutions of variable p_{CO_2} to calculate C_{CO_2} in (6.9).

2. **Uncertainty Quantification**: The characterization of stochasticity considered in this design and planning problem was already extensively discussed in 6.2.2., and the interactions among modelling components graphically captured by Figure 6.3. In this chapter, however, uncertainty on market-support instruments was also introduced, and simulated based on the values in Table 12. The resulting uncertainty space for the PCs defined is graphically captured by Figure 7.2 over 2000 *i.i.d* scenarios, showing both deterministic projections and stochastic realizations for each variable. The graphic confirms the need to implement a data-driven approach to find feasible design alternatives in such a high dimensional context. In fact, as implied by Figure 7.2, a dynamic and generalizable decision-making policy will be essential to capture the greatest long-term value out of the potential future governmental policy tools.

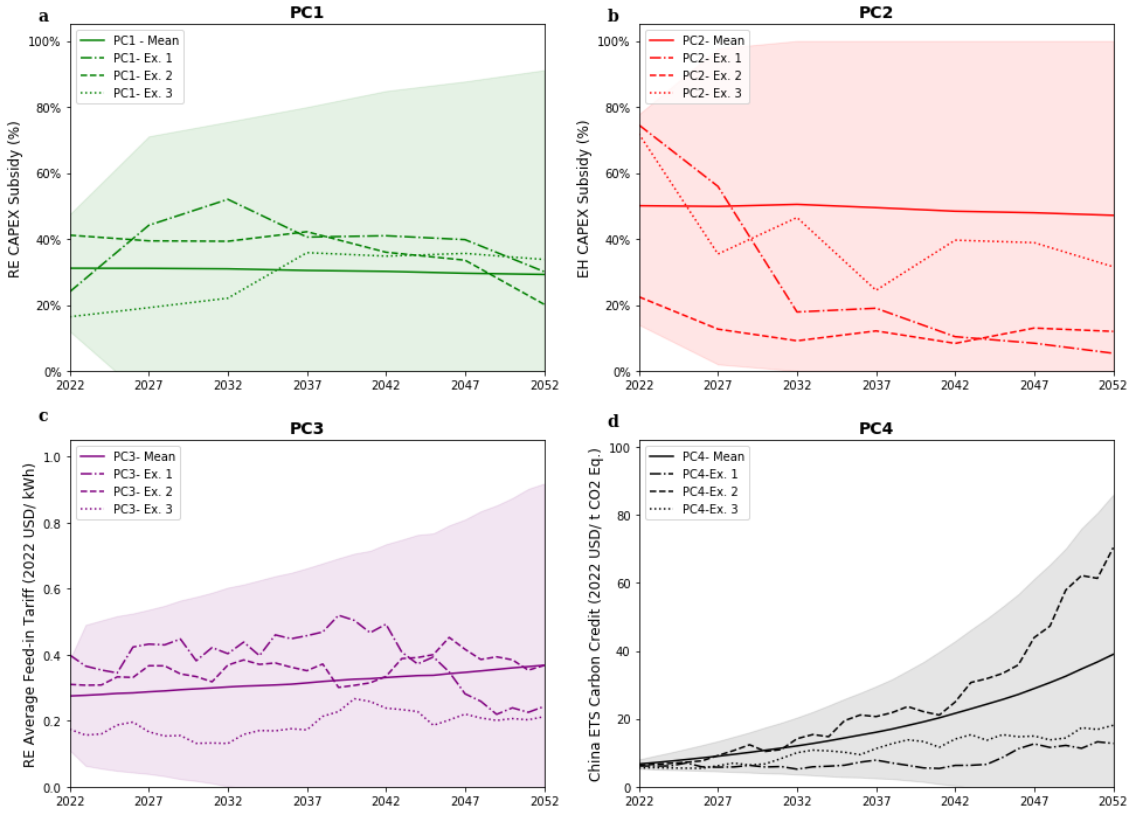


Figure 7.2: Uncertainty space considered for the different policy alternatives over the project 30-year horizon (a-d). PC1 is shown in (a), PC2 in (b), PC3 in (c) and PC4 in (d).Note that for each PC the shaded area represents the 95% confidence interval (assuming a normal distribution), the solid line the mean expected value over time, and the other lines 3 randomly selected example scenarios for each PC to highlight variability across GBM uncertainty realizations.

3. **Action Space Creation:** The decision-making possibilities for the agent are the same as those captured by (6.27). This is because energy system operators are not able to influence the state variables related to policy-making tools, thus it is appropriate to evaluate flexibility strategies based on the same technical enablers for mobile microgrids considered in 6.3.1.

7.3.2 Computational Framework for Data-driven Policy-making

In order to handle the increased dimensionality introduced by uncertain PC realization, a DRL based approach is also implemented to evaluate the performance different PCs, as captured by Figure 7.3. The generic iterative framework depicted in Figure 7.3a highlights the importance of starting by identifying the appropriate policy tools and constraints for a decision-makers objectives, It follows by recognizing their potential long-term uncertainty realization in the second stage, which points to the need to develop flexible and adaptable energy system solutions. Accordingly, the design and capacity adjustment strategy are then optimised through the DRL flexibility analysis approach depicted in Figure 7.3b. Cost-minimization objectives would be typically prioritized in this context, although other metrics can be easily subsisted, considering the entire uncertainty space captured by Figure 7.2. The respective performance distribution is used in the following step to generate important insights on the most effective PC parameters for energy-access metrics under uncertainty. Finally, their values and implementation can be refined through further stakeholder interactions and prioritised objectives in the final step. Nonetheless, the iterative nature of the framework naturally means these steps can be repeated several times to incrementally improve the formulation of market-support instruments under uncertainty.

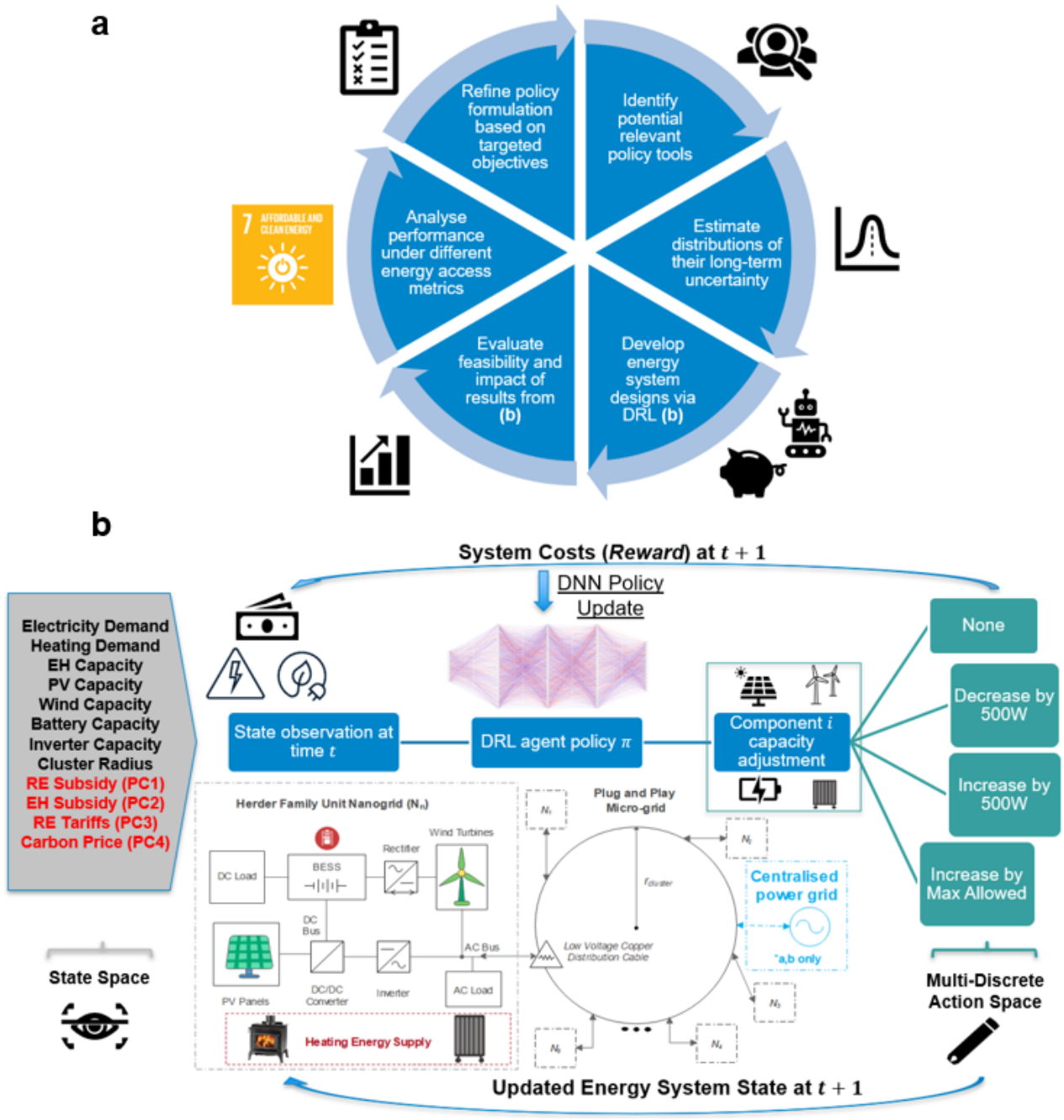


Figure 7.3: **(a)** Data-driven framework for enabling the evaluation and refinement of different policy alternatives to accelerate the transition to SDG7 under a high degree of uncertainty on energy system operation. **(b)** DRL agent iterative training process for energy system planning optimisation going from entire state space inputs (left) to selected action in a particular time step (right). The multi-energy vector block diagram shown on the bottom half of the figure represents, going from left to right, an example individual nano-grid unit, the PP nomadic community microgrid. Engagement with the centralised power grid is available only during the scenarios captured by Figure 7.1a and Figure 7.1b. The uncertainty in PCs introduced into the state vector is marked by the red variables (left). Note that, in fact, **(b)** is overall similar to Figure 6.8a, although the former highlights the increased state space considered in this Chapter through the inclusion of policy uncertainty. Additionally, while the heuristic for action node selection from Figure 6.8b is not shown here to enhance visual clarity, it is implemented exactly in the same manner during energy system operation.

7.3.3 Algorithmic Approach

The same algorithmic approach as in Chapter 6 was implemented as a starting point in this section since it already proved to be effective on the design and planning of flexible mobile microgrids. This also allows for validation and comparative evaluation with the results presented in 6.4. Please refer to 6.3.2 and Algorithm 3 for a more detailed overview. The primary difference between the environment models formulated in 6.3.1 and 7.3.1 is the novel PC state variables introduced in (7.1), as opposed to the original definition of the observation space in (6.25). Initial empirical results, however, suggested a substantial increase in complexity and variability for training the ACER agent with the very large PC uncertainty space implemented in this Chapter (graphically captured by Figure 7.2). This trend held across different hyperparameters and DNN architectures.

In light of this issue, several other algorithms (initially with default hyperparameters) were trained on the environment described in 7.3.1, using DRL packages designed for fast experimentation among different agents[166]. This analysis pointed to Proximal Policy Optimisation (PPO) as striking the best compromise between final performance, computational budget for training, and sample efficiency. The trained agent also yielded very similar flexibility strategy and results to those presented in 6.4, which uses PC0 assumptions from Table 12. Its logic and implementation are further discussed below.

PPO is based on many of the same principles as TRPO, focused on limiting policy update steps in order to avoid performance collapse over training[151]. Its method to so, however, is significantly simpler as it does not necessitate 2nd order approximations. PPO-Clip (hereafter simply referred to as PPO) is the most commonly used variant, which does not rely on KL-divergence measures, and exploits some of the advantages from actor-critic methods, such as an explicit value function representation. It implements stochastic policies to explore the decision-making space, although action selection converge over the course of training as its randomness progressively reduces. Furthermore, PPO can be highly parallelizable, as it exploits a vectorized architecture to concurrently collect trajectories from multiple environments.

Policy updates are limited by "clipping" in the objective function, aimed at reducing differences between new and old policies. The updates of the policy are generally performed according to (7.2). Multiple minibatch steps are typically used for maximizing the objective function L shown in (7.3). Here, ϵ is the clipping ratio parameter used to determine allowable divergence between consecutive policy updates. The advantage estimate, introduced in 2.3 is noted by $A^{\pi_{\theta_k}}$ here, and is approximated in this implementation through the GAE[150] method also used in Chapter 5. The clipping behavior is dictated by g, defined in (7.4). Case (i) increases the objective from (7.3) and corresponding probability of a selected action, while the reverse is applicable to case (ii). The total changes in objective function value are then limited by the min operator in (7.3). Additionally, the approximated KL-divergence D_{KL} is used in this chapter as an early stopping criterion (rather than the hard

constraint found in TRPO). In summary, therefore, PPO clipping is based on the recurring principle that there is limited value in performing excessively large policy (or value function) updates.

$$\theta_{k+1} = arg \max_{\theta} \mathbb{E}_{s,a\sim\pi_{\theta_k}}[\, L(s,a,\theta_k,\theta)] \tag{7.2}$$

$$L(s,a,\theta_k,\theta) = \min(\frac{\pi_\theta\,(a|s)}{\pi_{\theta_k}(a|s)}\, A^{\pi_{\theta k}}(s,a), \text{clip}\,(\epsilon, A^{\pi_{\theta k}}(s,a)) \tag{7.3}$$

$$\text{clip} = g\,(\epsilon, A) = \begin{cases} (1+\epsilon)A\ if\ A \geq 0 & (i) \\ (1-\epsilon)A\ if\ A < 0 & (ii) \end{cases} \tag{7.4}$$

The complete pseudocode for the PPO agent implemented for nomadic policy-making formulation is shown in Algorithm 4 below. As the shorter length compared to Algorithms 1-3 suggests, the code and optimization procedure are simpler to understand and implement than most alternatives. Additionally, it usually has very few hyperparameters to optimize, while achieving state of the art results in varied environments. Most recently, in fact, PPO has received renewed attention as the primary algorithm used for training the highly popular language model ChatGPT[72], which has performed exceedingly well across a variety of tasks. Based on the above, it is one of the most commonly implemented DRL algorithms in both research and industry applications.

Algorithm 4: Proximal Policy Optimization (PPO), with GAE, KL-based early stopping, and entropy bonus applied to the nomadic community PC analysis and formulation tackled in this chapter, adapted from[151].

Initialize: Input policy parameters θ_0, value function parameters $\emptyset_0$, clipping coefficient ϵ , divergence limit D_{KL}^{max}, Maximum iteration steps K, remaining hyperparameters (complete list given in 11.8)

for k = 0, 1, 2, … K **do:**

Generate stochastic evolution scenarios ${\xi_k}^i$ (including PC values captured by the state vector in (7.1)) according to (3.11)-(3.12), and model assumptions from Table 12.

Collect set of trajectories D_k by running policy $\pi_k\,(\theta_k)$ on the environment from 7.3.1

Compute rewards to go R_t through (6.26)

Estimate advantage $A^{\pi_{\theta k}}$ based on current value function $V_{\emptyset_k}$ using the GAE variant from (7.5)

Update the policy $\pi_{\theta_{k+1}}$ through the Adam[124] optimizer by maximizing the PPO objective in (7.3), with clipping change magnitude limited from (7.4).

Approximate difference between new and old policy $D_{KL}\,(\pi_{\theta_k}, \pi_{\theta_{k+1}})$ using (5.15)

12:	**if** $D_{KL}^{max} < D_{KL}(\pi_{\theta_k}, \pi_{\theta_{k+1}})$ **then:**
13:	**S**top gradient steps and do not update policy (early termination).
14:	**Else:**
15:	Refit value function $V_{\emptyset_{k+1}}$ through regression minimization of the mean squared error using
16:	the Stochastic Gradient Descent Momentum (SGDM) method [277], based on observed
17:	trajectories D_k under old policy π_{θ_k} and corresponding recorded rewards R_t
18:	**end for**

7.3.4 Training and Testing

Based on the above and previous empirical studies evaluating the performance of different PPO variants across different environments[320], a few minor changes are made from Algorithm 4 to enhance the learning process. They are briefly discussed here, although please refer to the respective publications for a more detailed overview. An entropy bonus α is included to incentivize additional exploration across the very large uncertainty space considered[321]. The objective function from (7.3) is also modified slightly by integrating an error term for value estimations. Aiming to address computational expense and memory limitations, minibatches are used for gradient computations, policy updates, and advantage normalization[125]. Rollouts were, accordingly, based on partial fixed-length trajectories of 60 steps (i.e. 5 years) in the environments. Learning rates are annealed throughout the training process, where the Adam[124] optimizer updates value function parameters, while SGDM[277] is used for policy updates. Finally, rewards are scaled by dividing them from the standard deviation of the cumulative rolling rewards, as opposed to the commonly used mean[320].

The PPO agent detailed in Algorithm 4 is trained using a budget of 72 Million monthly decision timesteps, equivalent to 200,000 simulated full 30-year episodes, allowing a more appropriate comparison to the results from 6.4. As performed on a standard laptop Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz, the complete learning process takes approximately 6 hours. The training environment assumed the potential for all PCs to materialise to some degree based on the parameters from Table 12. All training scenarios are stochastically generated to be *i.i.d* from each other, thus leading to an *infinite* number of potential evolutions and state-action pairs. This is particularly important here, as we seek to develop energy system designs and flexibility strategies which are generalizable across the entire policy-tool uncertainty space from Figure 7.2 (in addition to the general model stochasticity detailed in 6.2). As mentioned in the former section, the limited number of hyperparameters for PPO was optimised faster and more intuitively as compared to other DRL

agents. The final values used for the implementation detailed in 7.3.3 are determined across 100 experiments of 100,000 training steps using Optuna[170]. The complete list of the ones used in this Chapter can be found in 11.8.

Shared architecture and parameters are implemented for the DNNs capturing the policy and value functions, as previously shown to improve performance in higher dimensional environments [320]. These are akin to the actor and critic components described in 6.3.3, as shared network parameters presents a number of practical advantages. Simple fully connected MLP networks are thus implemented here, similar to the ones from Chapter 6, although using the tanh (hyperbolic tangent) activation function between hidden layers. Furthermore, one extra hidden layer (increasing the total to six) with 64 neurons is included to address the additional problem complexity introduced by PC uncertainty. Accordingly, the input layer is modified to integrate the PC variables captured by the state vector in (7.1), increasing observation space dimensionality from 13 to 17. The outputs from both MLP networks match the action space size from (6.27), as detailed in 6.3.3, and also use a SoftMax function to obtain a probability distribution over decision-making options.

The complete learning curve for the PPO agent on the environment detailed in 7.3.1 is captured by Figure 7.4, averaging results from 5 separate training runs using different random initialization seeds. Note that only the best training run is used for the performance analysis presented in 7.4. The evolution of reward observed during the training process suggests PPO to be an effective agent for this case study. Several key differences with the results for ACER depicted by Figure 6.9 in 6.3.3 can, however, be noticed. Firstly, the overall shapes of the curves confirms the better sample efficiency by ACER compared to PPO, as the majority of learning happens in the first 18 Million steps for the former, while being more distributed across the whole process for the latter. Generally, the objective improves by around 75% in the first 36 Million steps for PPO, with gradual increases up to around 54 Million steps, after which the policy seems to have approximately converged. The higher dimensionality of the state space here likely contributes to this trend, and also initially leads to much greater costs (i.e. lower episode rewards) for the nomadic energy system considered. Similarly, the increased uncertainty space from the integration of stochastic PCs results in an increased variance among PPO training runs. In essence, both these issues are driven by a necessarily deeper exploration of flexibility strategies to address the higher problem complexity found in this Chapter. Ultimately, average energy system costs obtained through the trained PPO policy are substantially lower than the ones from 6.4.2. Nonetheless, the episode reward values from Figure 6.9 and Figure 7.4 should not be directly compared due to differences in environment formulation. The PCs introduced in 7.2.4, in fact, can only contribute positively to lowering energy system costs for nomadic communities, regardless of their realization in different scenarios, giving some optimistic bias to the results presented.

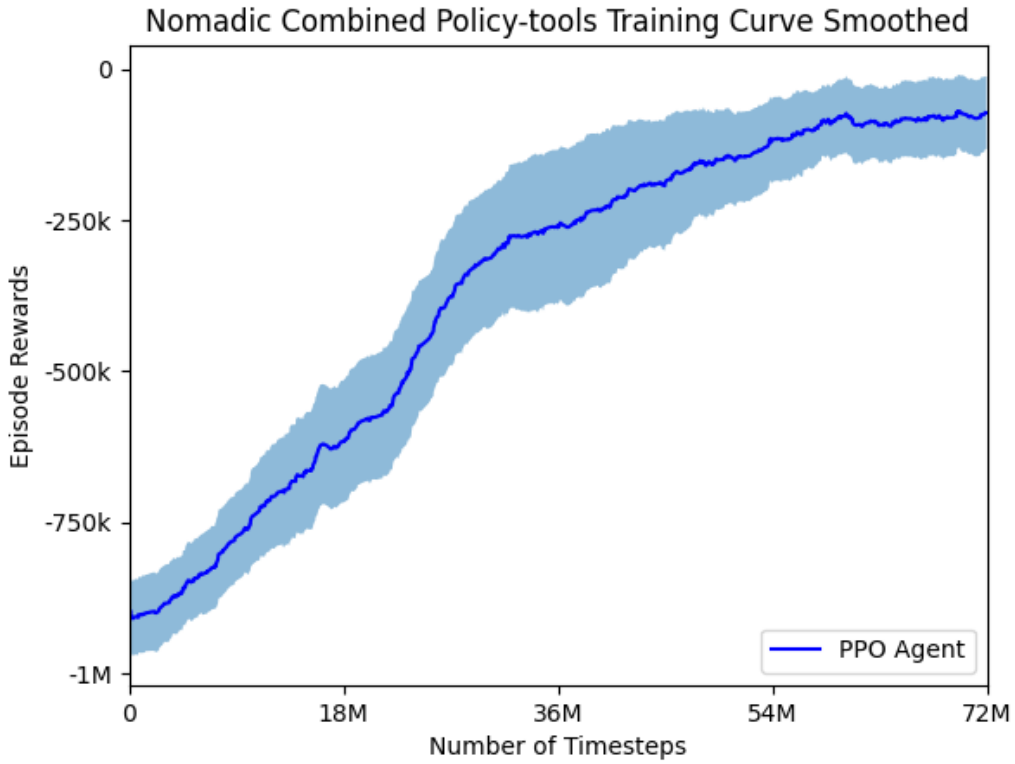


Figure 7.4: Training episode reward evolution for PPO agent on Mongolian energy system policy tool formulation environment. The y-axis estimated the total energy system cost per episode while the x-axis shows the number of *monthly* decision timesteps implemented for training. The solid line shows smoothed average (using a window size of 100 episodes) among 5 runs. The shaded area the standard deviation recorded across the same training runs.

The testing approach in this chapter is designed to be more comprehensive than the ones previously presented in this thesis, looking to quantitatively capture policy-making implications from both the perspectives of governmental decision-makers and nomadic communities. In order to address this, six separate testing environments (TE) are developed, as summarized by Table 13, decomposed by each potential policy formulation scenario. Accordingly, TE0 computes performance if no market support instruments are implemented over the 30-years considered, defining the baseline. On the other hand, TE1-TE4 assume the development of only a single policy-tool in all scenarios, altering the cost coefficients for technologies and key variables that the flexibility strategy was originally based on. In essence, this means TE0-TE4 evaluate performance with significantly different stochastic model parameters than trained for. Crucially, however, action selection in all of them is determined by the single generalised PPO agent policy trained as shown in Figure 7.4, based on the environment from 7.3.1 (captured by Figure 7.3). As such, most of the results shown in 7.4 constitute a novel *out-of-sample* testing approach, complementing the method initially implemented in 5.4.2. The marginal cumulative performance improvement for each PC is then estimated relative to the baseline (PC0) through the counterfactual effect, exactly as shown in the study by [316].

Table 13: Summary of testing environments used for nomadic policy-tool formulation objectives in this chapter.

Testing Environment (TE) Description	**Differences from 7.3.1 based on relevant state variables in (7.1). Applicable across project life T**
TE0: No applicable energy policies for nomadic communities (PC0)	Absence of PC1-PC4 (same assumptions as in 6.3.3) $RE_{subsidy}, EH_{subsidy}, RE_{feed\ in\ tariff}, p_{CO_2} = 0$
TE1: Materialization of PC1 uncertainty only (Figure 7.2a)	$RE_{subsidy} \neq 0$ (simulated based on row 1 from Table 12) $EH_{subsidy}, RE_{feed\ in\ tariff}, p_{CO_2} = 0$
TE2: Materialization of PC2 uncertainty only (Figure 7.2b)	$EH_{subsidy} \neq 0$ (simulated based on row 2 from Table 12) $RE_{subsidy}, RE_{feed\ in\ tariff}, p_{CO_2} = 0$
TE3: Materialization of PC3 uncertainty only (Figure 7.2c)	$RE_{feed\ in\ tariff} \neq 0$ (simulated based on row 3 from Table 12) $RE_{subsidy}, EH_{subsidy}, p_{CO_2} = 0$
TE4: Materialization of PC4 uncertainty only (Figure 7.2d)	$p_{CO_2} \neq 0$ (simulated based on row 4 from Table 12) $RE_{subsidy}, EH_{subsidy}, RE_{feed\ in\ tariff} = 0$
TE5: Materialization of full PC1-4 uncertainty (Figure 7.2 a-d)	None (simulated based on full Table 12 assumptions as in 7.3.1) $RE_{subsidy}, EH_{subsidy}, RE_{feed\ in\ tariff}, p_{CO_2} \neq 0$

The solution quality of the PPO policy is also computed through TE5, using the same stochastic model parameters that the agent was trained on. While perhaps less relevant to policy-making objectives, TE5 helps analyse DRL-based flexibility strategy performance under deep uncertainty on market support instruments, gaining some important insights for nomadic communities. Note that the reward penalty for budgetary and mobility constraints from (6.26) is removed for all testing, as in Chapter 6, evaluating the true objective function over 2000 simulated stochastic scenarios. Finally, it is reminded that action selection is determined solely on minimization of costs (defined in 6.2.1.7) during training. Nonetheless, the performance of system evolution strategies in TE0-TE5 are also broken down by their respective sustainability (defined in 0) , resilience (defined in 6.2.1.9), and energy access metrics (defined in 7.2.3).

The different time step resolutions implemented for PPO agent training and testing are summarised by Figure 7.5. It should be noted that, while not shown in the graphic, state space values for PC1-4 are assumed to evolve through a resolution of 1 or 5 years (dt parameters in Table 12), meaning they are only updated every 12 or 60 decision timesteps, respectively. As mentioned in 7.2.4, this is deemed more representative of the governmental policy-formulation process. Other timestep resolutions considered remain consistent across training and all testing. More detailed numerical results and analysis will be presented in the following section.



Figure 7.5: Summary of time-step resolutions implemented at different levels for DRL agent training and testing on the mobile microgrid model considered in this chapter.

7.4 Relative Performance Analysis

This section presents the results of mobile microgrid performance based on the policy-tools formulated in 7.3.1. No benchmarking with inflexible designs is discussed here given the DRL based solution was found to significantly outperform alternatives in 6.4, across all metrics evaluated. Additionally, the results from 6.4.7 show that the flexible design is much better suited to capture upside potential from policy realisations, further justifying the limited added value from such a comparison. This is particularly true given baseline designs do not consider capacity adjustment options. Accordingly, their stochastically optimal sizing is not relevant to the objectives of this Chapter, as all decisions on energy system configuration are made at project start. Consequently, they cannot be based on realizations of market support instruments over time, and would only be optimised on their full uncertainty range, with limited applicability in practice.

In summary, the testing process detailed in Table 13 aims to systematically decompose the *out-of-sample* performance of potential governmental nomadic energy policies under deep uncertainty through a data-driven approach. A DRL-Flexibility analysis formulation is used to produce the designs and strategies under different assumptions, with all decision-making based on a single and generalised DNN parametrized policy. Nonetheless, the relative performance of the more specialized ACER agent developed in Chapter 6 is also briefly discussed below, in relation to TE0.

7.4.1 Cost, Sustainability and Resilience

Even though this chapter is focused on achieving SDG7 for nomadic communities, it is essential to also evaluate the economic feasibility, sustainability, and resilience of optimised energy system

solutions, as these may be important limiting factors to their deployment in practice. They are assessed as in Chapter 6, using the ENPC detailed in 6.2.1.6, the lifetime carbon equivalent emissions detailed in 6.2.1.7, and the total unmet load detailed in 6.2.1.9. Their evaluation is captured in Figure 7.6**,** separated by each PC (or TE from Table 13). The impact of increased uncertainty is captured on each metric, as the baseline (PC0) consistently presents the narrowest performance distribution compared to all other PCs considered. Note again that PCs 1-4 represent renewable subsidies, electric heating subsidies, feed-in-tariffs, and carbon credits – respectively.

Figure 7.6a suggests that all PCs can lead to a reduction in expected net present costs from the baseline, with PC1 yielding the lowest average value, followed by PC2, PC3, and PC4, respectively. PCs 1-4 capture Interestingly, PC2 presents the best economic upside potential, achieving the lowest overall system costs in the tail of its distribution, as shown in Figure 7.6a. This is likely driven by strong heating demand growth and/or limited electricity load evolution in selected scenarios, as well as the larger uncertainty space considered compared to PC1. PC2 also performs best regarding system sustainability, both in terms of expected value and upside potential, as shown in Figure 7.6b. Notably, PC3 performs worse on average than PC0 on this metric, and overall, similarly to the baseline expected value on energy security, captured by Figure 7.6c.

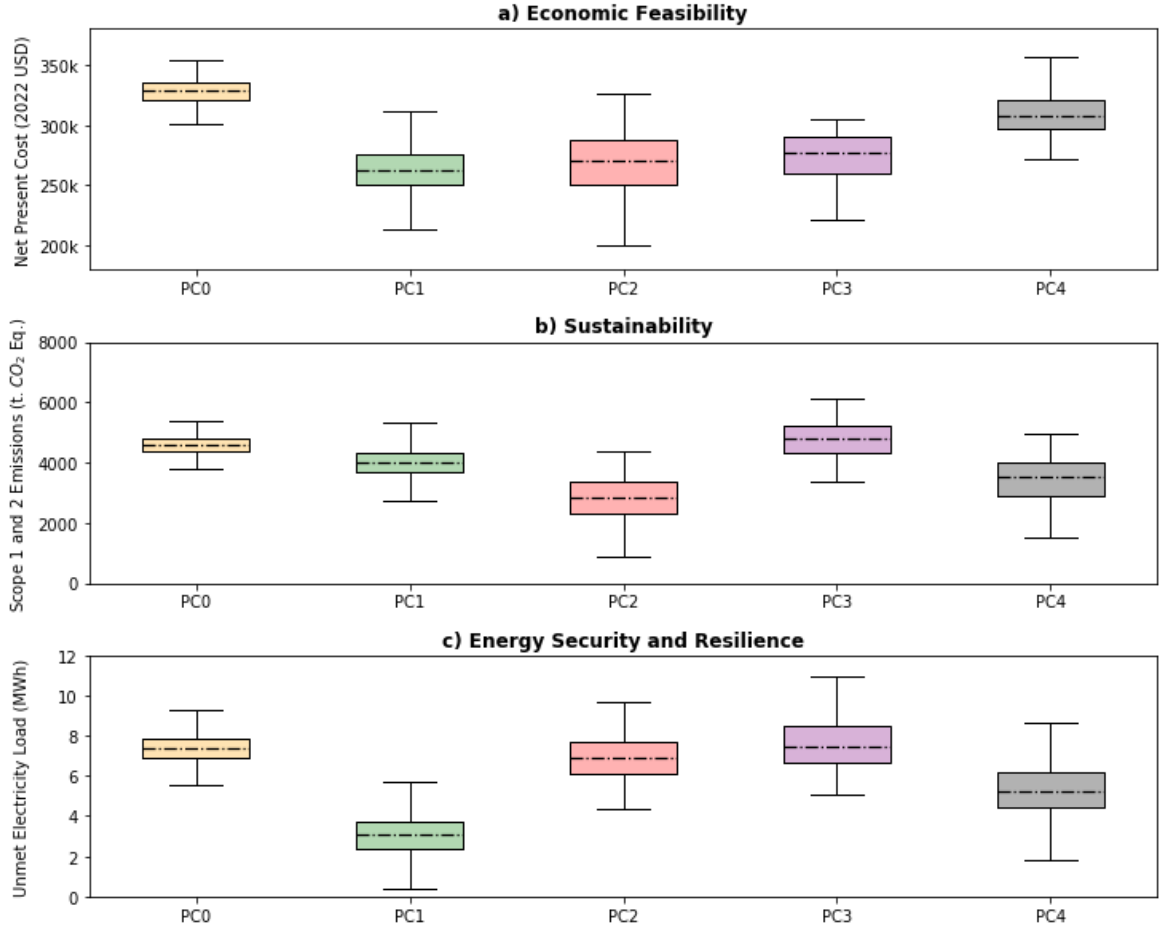


Figure 7.6: Multi-objective **(a-c)** evaluation of mobile and flexible energy system solutions deployed under different PCs (using their respective testing environments) capturing their expected performance under uncertainty outside of energy access metrics. The distribution of economic feasibility is captured by the boxplots in (**a**), system sustainability in (**b**), and energy security in (**c**). None that for **(a-c)** pictured above, the boxes represent the interquartile ranges under each PC for that metric, while the whiskers show the Value at Gain (**VaG**) and Value at Risk (**VaR**), or the values at the **95%** and **5%** percentiles of the performance distribution, respectively, over the testing scenarios analysed under uncertainty. The **mean** or expected value for each metric is shown by the **dotted black line in each box**, adopting the same colours as Figure 7.7 for consistency.

7.4.2 Energy Access Metrics

The energy access focus of this Chapter is captured by assessing the evolution of the most relevant SDG7 indicators for a typical 18-ger community mobile energy system, using the same measurement approach as the study by [316]. The results suggest that the PCs evaluated can significantly accelerate the transition to nomadic energy access across several metrics. The best performing policy tools, according to Figure 7.7, were PC1 for indicator 7.1.1*, PC2 for indicator 7.1.2*, PC1 for indicator 7.2, and PC3 for indicator 7.b.1. Over the study period, they generated impressive cumulative percentage improvements (i.e., counterfactual progress detailed in 7.3.4) of 34.9%, 47.8%, 23.5%, and 32.4%, respectively. The results for access to clean heating stabilize at a value of approximately 94% by 2032 for PC2, a great improvement compared to 81% several years later under PC0. The findings on share of renewable energy consumption are closest across PCs, converging quickly to values in the range of 90-95% (Figure 7.7c). This is intuitive given we only investigated renewable based systems, no dispatchable generators, and a limited load level, however, marks a significant improvement compared to current levels.

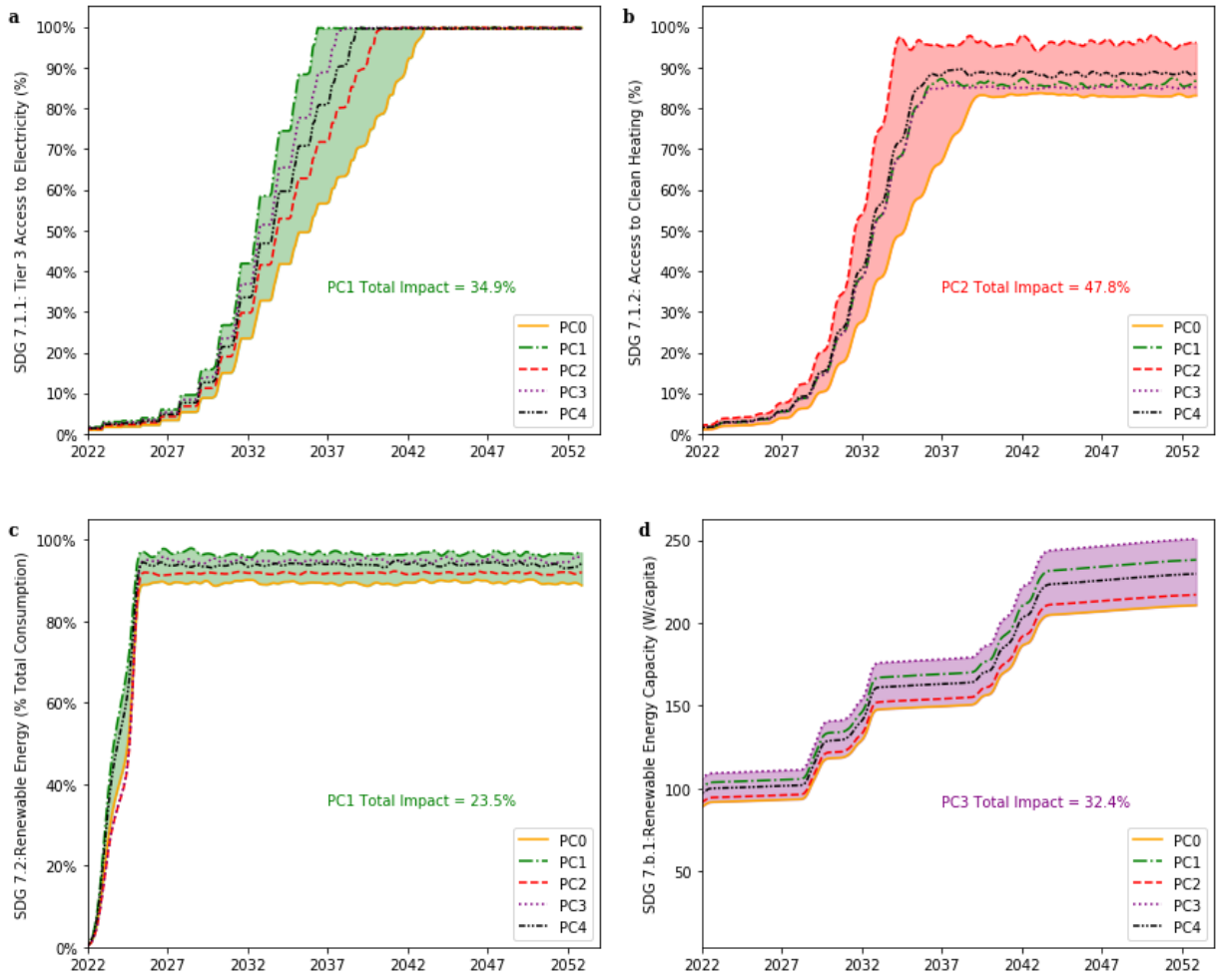


Figure 7.7: Performance of nomadic mobile energy supply system with the implementation of different PCs under uncertainty on energy access metrics SDG7.1.1 (**a**), 7.1.2 (**b**.) 7.2 (**c**) and 7.b.1 (**d**). The shaded area in between PC0 (orange line) and the best performing policy tool (stated) for each energy access metric (a)-(d) represents the cumulative percentage improvement generated by that PC.

7.5 Computational Insights for Policy-tool Formulation

The results from Figure 7.6 and Figure 7.7 provide some interesting initial quantitative insights to governmental or institutional policy-tool formulation objectives. The flexibility strategy developed by the PPO agents seems to be reflective of the budgetary constraints faced by global migratory communities, across all metrics evaluated in 7.4. More specifically, investment subsidies for renewables (PC1) are found to outperform alternatives in terms of economic feasibility (Figure 7.6a) and energy security/resilience (Figure 7.6c). Accordingly, PC1 is also most effective in term of energy access (Figure 7.7a) and share of renewables consumption (Figure 7.7c). In fact, it generates the greatest incentive to deploy renewable and BESS capacity within the microgrid, reducing reliance on the unstable main grid. Electric heating (PC2), on the other hand, is best suited to address heat supply decarbonisation and, in turn, access to clean heating (Figure 7.7b). The greatest policy impact was

indeed observed for EH subsidies, as energy systems in TE2 begin deploying more capacity earlier on in the project (Figure 7.7b).

Surprisingly, the highest renewable energy capacity per capita is achieved through feed-in-tariffs (PC3) (Figure 7.7d). PC3 seems to incentivize nomadic communities to deploy more RE capacity as they can sell excess generation at a premium, reducing total costs incurred for electrification in the long-term. A basic analysis of PPO action selection in TE3 suggests a tendency to deploy lower BESS capacities, prioritizing instead PP interactions with the main grid to capture potential profits from feed-in-tariffs in place. Accordingly, systems deployed under PC3 can lead to reduced overall energy security, driven by the uncertainty around grid availability, RE generation, and infrastructure limitations. PC3 also leads to higher overall emissions, as reduced storage levels tend to result in the need to purchase carbon intensive grid electricity or burn coal to meet heating demand.

Finally, the limited system sizes and low predicted prices of ETS make carbon credits (PC4) the typically least effective incentive considered in this study, particularly in terms of energy access metrics. This finding was not expected, given they are supposed to provide monetary incentives to deploy RE, EH and BESS capacity to maximize total avoided emissions. In certain high growth scenarios (i.e. cost > 2022 USD 55 by 2050), for instance, PC4 outperforms the alternatives on several metrics from 7.4. Therefore, while presenting more uncertainty on if and when it may actually be implemented, they may provide a more balanced approach to achieve the objectives targeted in this Chapter. This is particularly true if the market develops faster than currently projected, perhaps as a result of new emissions regulations.

Accordingly, the results for TE5 (defined in Table 13) are not shown here as they provide limited insights on the policy formulation problem, even if closest to the training environment from 7.3.1. Intuitively, they suggest that the best *average* performance can be achieved through a combination of PCs 1-4. This is true for all the metrics assessed in Figure 7.6, given the important synergies among them. For instance, RE subsidies incentivize capacity deployment as budget constraints become less restrictive. Similarly, feed-in-tariffs provide a revenue stream to recoup these investments when generation cannot be utilized. Combined, this can lead to a surplus in electricity production in many months, which may in turn be used to power EH and reduce coal burning to meet heating demand. Carbon credit revenues are then ultimately maximized based on the cumulative emissions reductions from above. Ultimately, this leads to the best economic feasibility, sustainability, and resilience of the mobile energy system considered across the testing scenarios evaluated. Nonetheless, the results for TE5 present a larger distribution of performance than when a single policy-tool is available, driven by the significantly greater combined uncertainty space guiding capacity adjustment decisions.

Nonetheless, a more advanced statistical decomposition of DRL-based energy system evolution under different conditions could be performed to improve triggering signals for policy-tool deployment. The

visualization techniques, decision rule generation, and design decision-support system detailed in 6.5, in fact, offer some important potential avenues for improved acceptability by decision-makers. While they were not performed in this Chapter to focus on the key policy-formulation contributions, the approach from 6.5.2 could be adapted to look at metrics more relevant to market support instrument implementation. These may include evaluating the distribution of capacity adjustment timings under each PC, or action selection magnitude based on their uncertainty realization. Many other options exist, and the analysis and selection process should be refined with relevant stakeholders. Ultimately, this can aid in generating insights on how the energy transition may be achieved (and enhanced) under each potential policy tool, or their combination.

Finally, the decomposed performance distributions obtained in the former sections suggest different tools may be effective depending on targeted objectives and the level of uncertainty observed. The similarity in the shapes of the curves shown in Figure 7.7 suggests a high level of generalizability for the DRL decision making policy under different realizations of uncertainty. Modular deployment strategies obtained are overall transferable across scenarios, although also dynamic enough to extract added value out of different PCs across the SDG7 metrics considered. This suggests the DRL agent can closely match generation capacity and demand, with the metrics for most PCs stabilising around the midpoint of the project as a result of the flexibility strategies implemented. Notably, Tier III access to electricity can be achieved under any PC over the next 16 years using the proposed DRL approach (Figure 7.7a). This result is very promising, suggesting that key SDG7 targets can indeed be met in Mongolia within a reasonable timeframe, although the transition to energy access is substantially accelerated in the presence of market support instruments. Furthermore, while this work only explicitly targeted SDG7, there are likely to be positive externalities on several other metrics in the Mongolian context. Their explicit assessment, however, was outside the scope of this Chapter.

7.6 Discussion

The analysis in 7.5 provided some general guidelines to evaluate which policy-tools tend to be most effective on the metrics considered here from a governmental perspective. The consistent overarching insight is that investment subsidies are generally very important for nomadic energy system policy-tool formulation, particularly in light of the low-income of these communities. The uncertainty on their actual implementation, however, must be reduced to capture the greatest value of them. Generally, in fact, episodes resulting in high system costs seem to be linked to the PPO agent expecting the materialization of some market support instruments in scenarios where this does not happen. As a result, excessive capacity may be deployed for various technologies, assuming investments will be recouped through the realization of one of the PCs considered here. Particularly valuable may be a detailed sensitivity analysis to stochastic (or even deterministic) model assumptions for key PC variables. This allows decision-makers to visualize the impact of implementing different regulations to those defined in Table 12. Extending this to all stochastic model parameters could

provide a deeper understanding on the impact of uncertainty and interactions among system components.

A comparison between the DRL agents developed in 6.3 and 7.3 confirms tensions between generalizability and expected performance through a data-driven approach. The generalised PPO policy performance is remarkable given the ACER agent training took around 33% longer in terms of wall-clock time, on the same number of training steps. Partly, this may be attributed to the improved parallelization capabilities of PPO, where trajectories D_k (line 6 from Algorithm 4) are collected using several vectorized environment initializations. Nonetheless, PPO resulted in slightly greater expected costs than ACER in the absence of any energy policies. This is somewhat intuitive given the ACER agent was more specialized to this testing environment, and consistent with trends observed in the out-of-sample testing from 5.4.2, or results from the literature on other environments (see 2.3). Interestingly, the generalised PPO agent assessed on TE5 (i.e. with the full policy-tool uncertainty space) outperforms ACER on most of the metrics considered in this Chapter. While somewhat unintuitive, this is driven by all policy-tools benefitting the energy transition. Under deep uncertainty, however, this leads to a much broader performance distribution for PPO as compared to ACER, which may introduce some concerns on acceptability. Accordingly, it is also possible to train one agent per PC and likely improve specialized performance, but this kind of misses the motivation of this chapter to capture the full impact of uncertainty on market support instruments in energy systems designs. If more updated or certain information on actual PC development was available, however, this approach could extract additional value out of flexibility strategies for nomadic energy policies.

The results from 7.4 and corresponding insights described in 7.5 and this section, it should be noted, are completely contingent on implementing a DRL Flexibility Analysis approach to the planning of future nomadic energy systems. As such, an important limitation of this work is that the targeted nomads are assumed to completely follow the DRL agent suggestions for their long-term energy system evolution. This may be unrealistic in light of difficult to quantify cultural or technological factors, which may limit the adoption of the proposed approach in these highly traditional communities.

Another important factor to consider in extensions of this work is the total cost incurred for implementing each PC, evaluating the *cost-effectiveness* of achieving specific objectives. Additionally, it is still not clear how exactly these costs would be financed by the parties involved. A basic evaluation provided total costs of electrification for gers, although this was simply computed by extrapolating the results from the 18-ger system considered to the entire nomadic community in Mongolia. On the other hand, the costs for PCs 1-4 were also estimated by assuming they were completely covered by the government. This estimation, however, was based on the uncertainty space considered on the actual magnitude of market support instruments. In practice, the expense of

implementing such programs could be explicitly calculated as set values for them would be held constant over a pre-determined time period. Based on limited available data, therefore, a more advanced analysis could not be included at the time of writing, although it could further align this chapter with practical policy-formulation objectives.

7.7 Chapter Summary and Conclusions

This chapter is based on an extension of the policy-tool sensitivity analysis for mobile microgrid systems in Mongolia introduced in 6.4.7. The contributions of this work are exemplified in several ways. As a starting point, a novel computational framework is developed and implemented on the case study of ger energy systems planning. As such, a very large PC uncertainty space was integrated into the DRL-Flexibility Analysis formulation, with values changing significantly over project life.. A highly generalised PPO agent was then trained on this environment, developing energy system evolution strategies applicable across extremely varied scenarios. Additionally, a novel and more comprehensive testing approach was developed to decompose the combined policy-tool performance from the individual one. In summary, the results showed that all Mongolian SDG7 targets can be met over the next 30-years, but only assuming energy system capacity adjustment decisions are made through the proposed DRL-based approach. This is true under all PCs, which also show very promising results on economic feasibility, sustainability, and energy security/resilience. The results urge the importance of developing investment subsidies or carbon credits for financial and sustainability objectives, while deploying feed-in tariffs for increased energy security and main grid network stability. Most importantly, planning should be targeted at the community rather than single household level, and decision-support systems used to inform mobility choices. A generic overview of associated computational insights for policy-formulation was presented, although the same blueprint and process as shown in Chapter 6 may also be implemented to further improve understanding of DRL decision-making process in this context. Finally, this chapter presented an extensive discussion on the performance of a specialized agent compared to a generalizable one.

The completely different perspective considered here as compared to Chapters 3-6 suggest that there may be more applications of the approach developed in this thesis than originally thought. Essentially, this chapter marks a shift from pure engineering systems design and planning to actual governmental or institutional policy-tool formulation objectives. Other potential applications may be to inform electricity network development or other centralised infrastructure investments. While not components of the mobile microgrids targeted here, these factors can directly affect their feasibility and performance through several synergies and interactions. This is highly relevant to global migratory communities, where the many sources of uncertainty related to their energy access have thus far limited a quantitative analysis on the impact of different institutional programs. In fact, this

work diverges substantially from most previous studies on energy policy, which tend to use deterministic projections, or a few select scenarios, in their analysis of market support instruments.

Based on these issues, it is essential to holistically consider the trade-offs involved between a generalized and specialized DRL-based flexibility strategy, and respective policy-formulation implications. Additionally, benchmarking these results compared to baseline rigid designs is likely to further enhance understanding on the added value introduced by the methodology proposed in this thesis. Overall, a generalized formulation seems more relevant and appropriate given the deep uncertainty on policy-making across the world, however this may vary depending on the application context. Furthermore, this is likely to result in a broader distribution of performance, which is not well suited to risk-averse decision-makers. In light of this issue, in fact, a more systematic approach to integrate risk-profiles into a DRL-flexibility analysis should be explored, with important implications in regard to acceptability.

Chapter 8

8 Risk-profile Integration

This chapter focuses on a key recurring issue for the acceptability of the DRL Flexibility Analysis approach presented in this thesis, the integration of decision-makers risk profiles. It is, in essence, based on the following question:

- RQ8: How can risk profiles be integrated through a DRL formulation in a more dynamic and systematic way than currently possible with alternative methods?

This is particularly relevant due to the "black-box" nature of the optimisation procedure involved, making it less likely for system operators to implement recommended flexibility strategies directly.

The above issue motivate the work in this chapter, as we look to essentially produce narrower performance profiles, or optimise flexibility strategies from a risk-averse or risk-prone perspective (compared to the risk-neutral one taken thus far). The high-level goal is to ultimately create more certainty and acceptability around the results and outputs from DRL applications in engineering systems. Developing appropriate metrics to estimate risk levels under uncertainty is essential to this, particularly ones compatible with the policy optimisation procedures used. More specifically, conditional value at risk (CVaR) is deemed most appropriate in this context, as further detailed below. Distributional DRL is found to be the better approach to optimise this metric, and is applied to the case study from Chapters 4. Results are compared to their respective risk-neutral implementations, showing significant value for this approach, and computational expense considerations are discussed.

Alternative methods based on the objectives of this chapter were also explored through a research placement in the Institute for Construction and Infrastructure Management at ETH Zurich, under the supervision of Prof. Dr. Bryan T. Adey, and in primary collaboration with Orlando Garcia. This research group is particularly well suited to complement the methodology developed in this thesis, as they focus on improving the effectiveness and efficiency of infrastructure management process. Logically, an important part of this is estimating risk and infrastructure performance under extreme events, looking for ways to embed resilience into the system to easily restore operation. While identified as an important consideration in earlier chapters, it has not yet been directly addressed in this thesis. Many approaches were explored together for this purpose. Ultimately, scenario clustering based on risk level provided important visualization capabilities to decision-makers. Accordingly, it was used to train separate DRL policies based on the perceived risk-level, helping to address the chapter objectives. This process was found to introduce additional trade-offs between generalizability and performance in select scenarios. Unfortunately, due to personal reasons, the placement at ETH

ended up being much shorter than planned. As a result, the proposed management technique is still under development, and does not fit directly within the scope of this thesis. Nonetheless, the placement still led to a collaboration on a journal paper presenting the proposed approach on a water infrastructure management case study (not here). It is currently in the final stages of editing and will be submitted shortly to the Journal of Infrastructure Asset Management[322].

The full open source code to reproduce all results presented in this chapter can be accessed here[217].

8.1 Introduction

Resilience in the context of engineering systems design and planning can be conceptualized as a means to respond to realizations of different risks [323]. The question of how best to address it has been studied extensively, leading to varied definitions, frameworks, and measures in the literature[37,324-327] – making it difficult to determine the most appropriate ones for a specific application. The integration of risk for the Flexibility in Design context considered here, in fact, seeks to capture a different objective than for typical resilience engineering. This is usually based on reducing downside exposure under uncertainty, as opposed to quickly restoring system functionality after disruptions. An important commonality among them, however, is the inclusion of a sequential decision making strategy or framework to be used during system operation as uncertain variables evolve.

Previous attempts to handle uncertainties and risks in resilience engineering have been typically based on chance-constrained optimization[328], robust optimization[329], and variants of stochastic programming[330]. The latter is the most commonly found[212,331], as it tends to be easier to solve and implement, while able to generate solutions tailored to the desired risk tolerance[332]. Nonetheless, these methods present many of the same limitations as standard ones in Flexibility in Design, or engineering systems planning under uncertainty – even if considering a risk-neutral objective[53]. For example, they may rely on function linearization, path independence assumptions, a small number of decision stages, or limited uncertainty recognition to obtain a tractable optimisation procedure. These limitations and the motivations for exploring a data-driven methodology were further detailed in 2.1.

The DRL based approach presented in this thesis addresses many of these limitations, as shown by the results from Chapters 4-7, but in its traditional form only seeks to optimize a risk-neutral objective. As such, the work presented here builds upon the approaches developed for improved interpretability of outputs in the previous Chapters (detailed in 4.5, 5.5, 6.5 and 7.5). Notably, some basic qualitative approaches or heuristics were already introduced for the objectives of this Chapter in these sections. For instance, these included ranking potential actions in terms of their overall investment risk to refine DRL agent recommendations (as in 4.5.2). Alternatively, when using a stochastic policy, the differences in action selection probability resulting from each state observation provide a more quantitative approach to account for risk revels. A threshold could be set (or optimised) to determine potential deviations from the DRL policy, and combined with a rank-order of expected investment

risk per decision for final selection by system operators (as in 5.5.2). Finally, the decision-rule generation from 6.5.2 provides novel and intuitive ways to estimate risk in action selection. For instance, this could be based on distributions of expansion magnitude and timing under different shortage levels or realizations of major uncertain variables, among others.

These initial approaches are a starting point to analyse risk in engineering systems design and planning. They fail, however, to provide a data-driven and systematic method to dynamically integrate changing risk preferences over time. Crucially, in fact, all DRL policies developed in the former chapters were optimised for expected value, or simply the average performance. While the VaR and VaG of solutions were evaluated, no implementation actually focused on improving performance on the tail-ends its distribution. This point is especially important given the general trend observed in methodological validation and extension stages of this thesis. Those results showed that while the DRL approach outperformed alternative state of the art methods in terms of expected value, its inherently greater decision-making flexibility yielded a broader distribution on evaluated metrics. The fact that VaR was worse for the DRL based solution than with other Flexibility in Design approaches in certain cases may be concerning in terms of applicability.

In fact, most of the historical algorithmic improvements in DRL have focused on maximizing mean cumulative rewards across environment interactions. This approach is valid under most settings, as a policy can be developed to determine decision-making under uncertainty through reward maximization over a number of scenarios. Only modelling the expectation of return, however, fails to properly account for agent performance in certain low-probability but high impact scenarios. Similarly, it does not consider the wide range of variability in return of a selected action. These are key concepts when considering risk-tolerance integration, as some decision-makers may prioritize worst case scenarios over the expected value across all uncertainty realizations.

Returns and their likelihood of occurrence could instead be characterized through a *probability distribution* in DRL given enough samples. The use of a return distribution was first popularised by [142], and motivated by the premise that it may provide a much deeper understanding of optimal agent decision making as opposed to the single scalar value normally computed. This can be thought of as a distributional perspective on DRL, exploiting information from the entire return distribution function. It helps address non-stationary policies in the environment, and capture complex underlying phenomena which may not be apparent through the state observation formulated[333]. As such, a return distribution is very useful to understand the aleatoric uncertainty (defined in 1.1) of environment interactions, which is present in all the case studies evaluated in this thesis. Additionally, the use of a return function captures the multimodality of value distributions, providing a broader spectrum of insights as opposed to a scalar approximation, and improving stability during the learning process.

Based on the above, distributional DRL has led to state of the art results on several benchmark tasks through a risk-neutral objective in the past few years [334]. The majority of studies, however, have focused on developing risk-adjusted agents which exploit the full information available from a distributional perspective. As a result of the probability function, decision-making policies can be optimised based on selected parts of the distribution, depending on their respective risk-tolerance. This approach to develop risk-sensitive policies has already been shown to be effective on varied environments, ranging from self-driving cars[335], to managing electricity usage in a building[336], to drone navigation[337] - among others.

Based on the above, the advantages of DRL-Flexibility analysis presented in this thesis can be extended to the integration of risk-profiles for different engineering systems, more dynamically, adaptively, and systematically than with alternative methods.

8.2 Background and Formulation

The core theory behind DRL is essentially that an agent should select actions based on environment observations with the aim to maximize its expected *return* following a decision-making policy π. As detailed in 2.3.1, this value is traditionally captured through the Bellman equation (i.e. $Q^{\pi}(s,a)$ in (2.4)). The random return in expectation $Z^{\pi}(s,a)$ however, may also be expressed by the recursive *distributional Bellman* equation, as shown in (8.1). This is also referred to as the *value distribution*, analogous to the value function in classical RL[333]. The difference between the two is also graphically captured by Figure 8.1, showing how Q^{π} outputs a single scalar value return, as opposed to the probability distribution returned from a state-action pair in Z^{π}.

$$Z^{\pi}(s,a) \overset{D}{=} R(s,a) + \gamma Z(s',a') \tag{8.1}$$

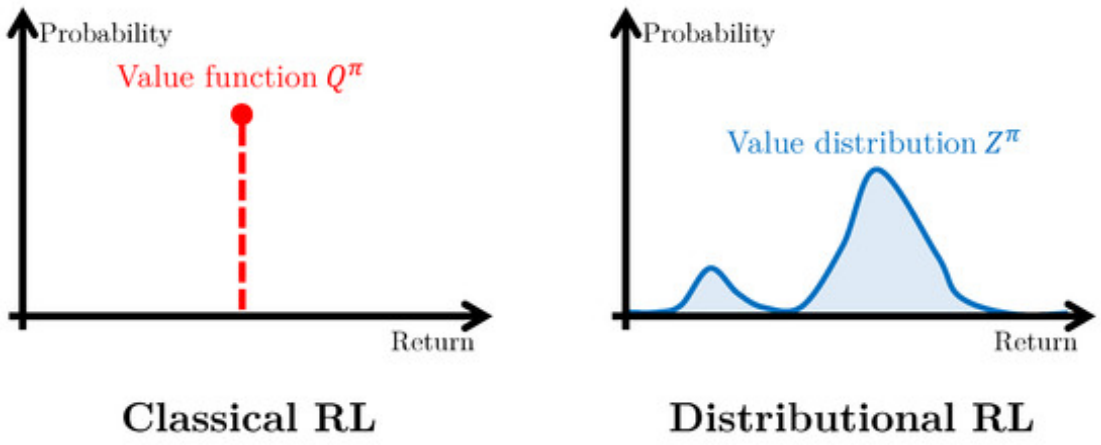


Figure 8.1: Visualization of difference between value function in classical RL and value distribution in distributional RL

Generally, distributional RL algorithms can be characterized by two distinctive features from classical RL. Firstly, the parametrization approach used to approximately capture the random return distribution, which typically is still within a DNN. Secondly, the distance metric (or loss function) between two probability distributions for improving approximations over the learning process. These are fundamentally based on the assumptions made about random returns and the granularity in approximations required by the problem.

Categorical DQN, also referred to as C51, employs a discretized categorical distribution (with 51 equidistant *atoms*) to parametrize returns and minimizes the KL-divergence to the target distributional Bellman projection [142]. As such, it assumes returns are bounded within a range known a priori, often leading to overestimation of variance as a result. Nonetheless, it outperformed all previous DQN improvements on the Atari suite. Following this initial work, QR-DQN was proposed to capture return distributions when they are not able to be fixed in a known range[338]. This algorithm implements quantile regression to automatically adjust a discrete set of return quantiles by minimizing the Wasserstein distance during value distribution updates. Implicit Quantile Networks (IQN) extends these work by learning the full continuous quantile function, mapping probabilities to returns[339]. During training, it reparametrizes samples from a base distribution τ (typically $\sim U([0,1])$) to the target distribution. This presents several advantages to earlier distributional RL agents, such as the ability to control the approximation error, improved sample efficiency, stability, and final performance.

Most importantly to the objectives of this chapter, however, it introduces the potential to formulate risk-sensitive policies based on the estimated return distribution. IQN implements distortion risk-measures based on expected utility theory, essentially shifting the cumulative probabilities of random returns and accordingly learning a decision-making policy through an altered expectation[340]. Based a distortion risk measure β:[0,1], where 1 is equivalent to risk neutrality, the distorted expectation $Q^{\pi}{}_{\beta}$ of Z^{π} under said measure is defined according to (8.2). An approximated risk-sensitive policy can then be found for K samples according to (8.3). Several distortion risk measures may be used to develop risk adjusted policies, however conditional value at risk (CVaR) is most intuitive and easiest to implement in practice[341]. CVaR has been extensively studied in the literature, particularly in finance applications, where it is typically referred to as the expected shortfall. In essence, it changes the distribution $\tau \sim U([0,1])$ to $\tau \sim U([0, \alpha\ ])$ for IQN sampling. As shown in (8.4), it represents the expected value of quantiles less than α, which can accordingly be used for action selection and policy formulation. It can thus be used to differentiate between standard and risk-sensitive DRL as shown in Figure 8.2. Intuitively, CVaR is the expected performance beyond an α cutoff point or confidence level, where the closer α is to 0 the more focused on the tail end of the distribution (i.e. risk-sensitive), while 1 represents risk-neutrality.

$$Q^{\pi}{}_{\beta}(s,a) := \mathbb{E}_{\tau\sim U([0,1]}[Z^{\pi}{}_{\beta(\tau)}(s,a)] \tag{8.2}$$

$$\pi_{\beta}(s) = \arg\max_{a\in A} \frac{1}{K}\sum_{k=1}^{K} Z_{\beta(\tau_k)}(s,a) \tag{8.3}$$

$$CVaR_{\alpha}\ (Z^{\pi}(s,a)) = \mathbb{E}_{\tau\sim U([0,\alpha]}[Z^{\pi}{}_{(\tau)}(s,a)] \tag{8.4}$$

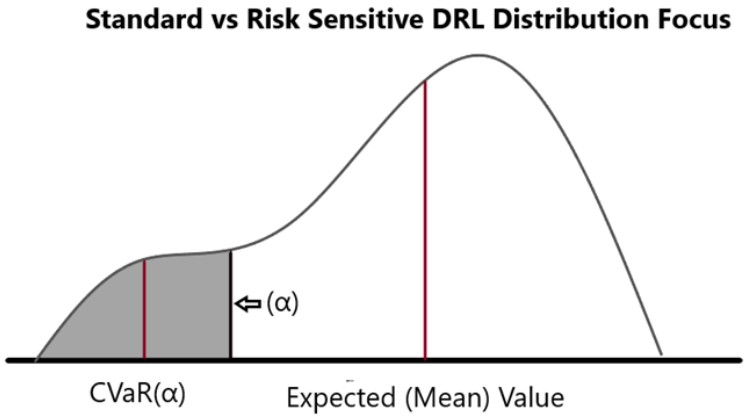


Figure 8.2: Comparison of distribution focus for standard vs risk sensitive DRL, with shaded region representing focus and return performance metrics used in risk sensitive DRL.

The IQN approach for estimating return distributions through sampling for quantile functions gives a rich amount of information and allows the development of risk-aware policies in several creative ways. This is a very new field and thus is constantly evolving, with a very up to date review of potential algorithms available at [333]. Nonetheless, one approach which is particularly interesting for the objectives targeted in this chapter is Ensemble Quantile Networks (EQN)[342]. The EQN agent provides an estimate of both the epistemic σ_e and aleatoric σ_a uncertainty from a specific state observation, based on the variance of returns and Q-values, respectively. The former exploits the distribution obtained through IQN directly, while the latter trains an ensemble of DNNs in parallel on different subsets of available data generated through a randomized prior function. Ultimately, this results in the development of a piecewise policy as shown in (8.5). Accordingly, EQN allows decision makers to determine whether to use the value maximizing policy if the risk level is deemed overall low (i.e. confident), or to switch to a backup or safe (i.e. more risk-sensitive policy) otherwise. The thresholds for this confidence level can be adjusted through stakeholder interactions based on project specific features. Notably, however, EQN was developed for autonomous driving applications where the development of a backup policy is more intuitive (i.e. stop before the intersection if possible).

$$\pi_{EQN}(s) = \begin{cases} {\pi_{IQN}}^{\alpha=1}(s) & if\ confident\ (i.e.\ \sigma_a, \sigma_e < treshold\ ) \\ \pi_{Backup}(s)\ \left(i.e.\ {\pi_{IQN}}^{\alpha<1}(s)\right) & otherwise \end{cases} \tag{8.5}$$

8.3 Distributional DRL Implementation

This section outline the development of risk-sensitive agents and compares the distribution of performance to a risk-neutral one. Results are only shown for the proof-of-concept case study initially tackled in Chapter 4 as it provides some better intuition on the trade-offs involved, which are harder to capture and break down on higher dimensional environments with many variables. Accordingly, the environment formulation that the results in this section are based on can be found in 4.3.1. Note that while risk-profile integration is only shown for the garage case study, the same method could be easily extended to the other environments – although this was outside the scope of this thesis. This serves to introduce how a data-driven perspective can allow for more dynamic and systematic risk-profile integration than with standard methods in Flexibility in Design. Further potential improvements to these results are discussed later in this chapter.

8.3.1 Algorithmic Approach

Although many agents are being continuously developed with risk-sensitivity in mind, the IQN method[339] is applied here as it is the foundational one which following ones have been built upon. IQN approximates within a DNN the quantile function $F_Z^{-1}(\tau) := Z_\tau$ of the random returns Z shown in (8.1). During training, networks weights are updated to improve the estimation through quantile regression, normally using two samples τ and τ' - sampled N and N' times respectively. The distributional TD-error $\delta_t^{\tau,\tau'}$ can thus be calculated as shown in (8.6), where a_{t+1} is determined through the sample based estimate of the policy according to (8.3). The quantile Huber regression loss ρ_k for the same quantiles is then found as shown in (8.7) using threshold k, where $\mathcal{L}_k$ is the *generic* Huber loss defined in [338]. Finally, using a minibatch M of sampled experiences, the complete loss function for IQN is computed according to (8.8).

$$\delta_t^{\tau,\tau'} = r_t + \gamma Z_{\tau'}(s_{t+1}, a_{t+1}) - Z_\tau(s_t, a_t) \tag{8.6}$$

$$\rho_k\left(\delta_t^{\tau,\tau'}\right) = |\tau - \mathbb{I}[\delta_t^{\tau,\tau'} < 0]| \frac{\mathcal{L}_k(\delta_t^{\tau,\tau'})}{k} \tag{8.7}$$

$$L_{IQN}(\theta) = \mathbb{E}_M\left[\frac{1}{N'}\sum_{i=1}^{N}\sum_{j=1}^{N'}\rho_k\left(\delta_t^{\tau_i,\tau'_j}\right)\right] \tag{8.8}$$

DNNs to capture the entire approximate return distribution for each action from each state are fully connected MLPs with a cosine embedding, ReLU activation function, and two hidden layers with 64 neurons each. This was reported to achieve the best performance on the Atari domain in the initial publication, and while some attempts were made to modify this architecture they were not recorded to yield any significant improvements. The logical progression of the agent is very similar to that

depicted in Figure 4.4, with the primary difference being the substitution of distribution Z^{π} as opposed to scalar values from Q^{π}, as depicted in Figure 8.1.

The simplified overview of the full training process is given in the pseudocode from Algorithm 5, as applied to the garage case study from Chapter 4.

Algorithm 5: Implicit Quantile Networks (IQN) training process to develop risk sensitive policies applied to infrastructure case study and environment detailed in Chapter 4, adapted from[339].

Initialize: Z network parameters θ, risk-sensitivity CVaR coefficient α, remaining hyperparameters (full list given in 11.9), Y training episodes

for Y episodes **do:**

    Initialize state observation randomly and select starting number of floors (f_0) based on initial policy

    for $t = 1$ to T **do**:

        Sample $\tau_1 \ldots \tau_K \sim U([0, \alpha\,])$ based on desired risk-sensitivity level distortion measure

        Determine action according to (8.3) or random ones depending on ϵ-greedy sampling.

        Collect set of trajectories m by running policy π_{CVaR} using α on the environment from 4.3.1

        Combine trajectories into experience minibatch M

        Update distributional network θ with SGD minimization of L_{IQN} from (8.8)

    End for

End for

8.3.2 Training and Testing

In order to gain a comprehensive understanding of IQN agent behaviour, it is implemented for three different levels of risk sensitivity. This means training was conducted separately for $\alpha = [0.25, 0.5, 1.0]$ respectively, where the latter is equivalent to a risk-neutral (RN) agent. Initially, even more risk-sensitive agents were explored (i.e. $\alpha < 0.1$), however empirical results found very limited exploration of the decision-making space at these values. Accordingly, this resulted in very low ENPVs for the environment simulations, comparable to those obtained with an inflexible baseline in 4.4. This suggested that for such high levels of risk-aversion a modular deployment approach did not seek to capture sufficient upside potential, so more extensive training was conducted using the more balanced agents.

The IQN agents described in the former section are trained using a budget of 200,000 yearly decision timesteps each, equivalent to 10,000 simulated full 20-year episodes. As performed on a standard laptop Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz, the complete training process for a single agent takes approximately 2.5 hours. The computational expense of IQN creates an interesting trade-off with improved sample efficiency for learning a policy compared to DQN. As such, a lower computational budget was used than in Chapter 4, with the relative performance further discussed in the following sections. Based on the same issues, none of the DQN augmentations used in 4.3.2 such as prioritized experience replay, or dueling networks are implemented here. Limited hyperparameters are optimised through Optuna using 20 experiments over 15,000 timesteps.

The complete learning curve for all three levels of risk-tolerance considered are shown in Figure 8.3. They show an overall similar shape and reward evolution across risk-sensitivities evaluated, implying the learning process is overall similar for the various IQN agents. As expected, however, the risk-neutral agent achieves the best expected performance, followed by the medium risk aware agent (α = 0.5) and the highly risk-sensitive one (α = 0.25) respectively. The graphic also shows a lower standard deviation among runs as risk-sensitivity increases, which can be attributed to the more limited exploration of the environment decision-making space.

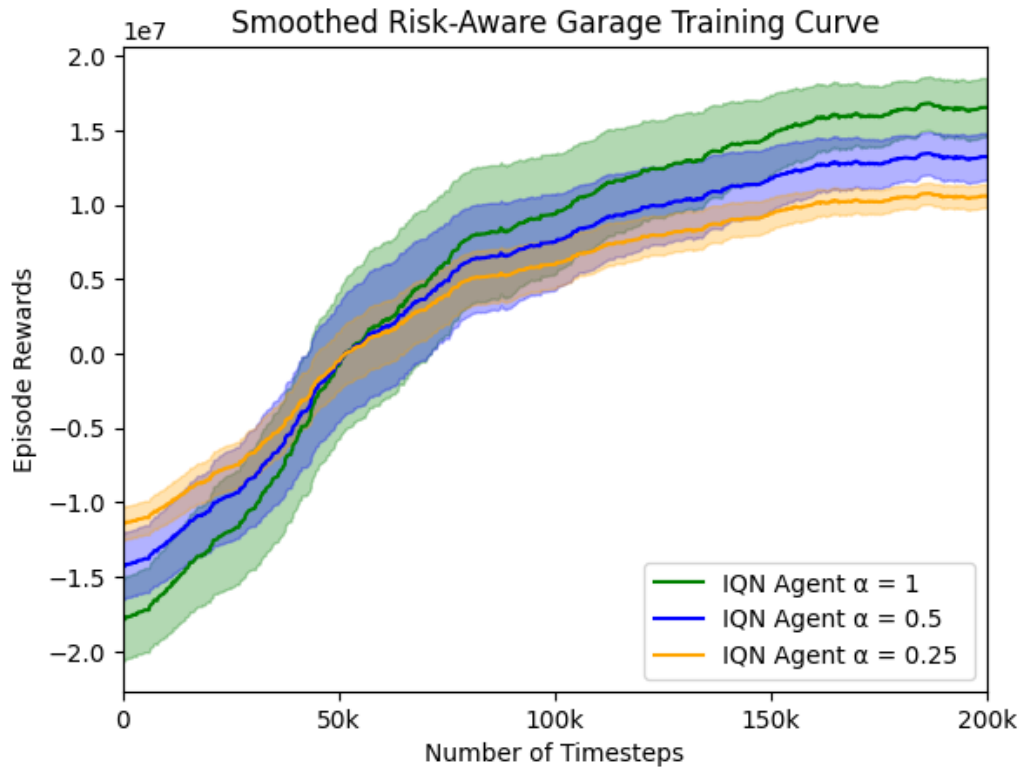


Figure 8.3: Learning curve for IQN agent on garage case study using three different levels of risk sensitivity.

The training and testing process is otherwise as detailed in 4.3.3, and the timestep resolutions used in this Chapter are summarised by Figure 8.4.



Figure 8.4: Summary of timestep resolutions across training and testing process used for risk-profile integration on garage case study.

It should also be noted that the EQN approach, while promising from a theoretical and decision-making intuition standpoint, was found to be excessive in terms of computational expense – hence why it was not further pursued or presented here. In fact, training for the same number of timesteps was found to take around 10-13 times the wall-clock period taken for IQN agents learning. This is due to having to train several DNN ensembles in parallel to estimate epistemic and aleatoric uncertainty, which while interesting, are not really necessary to develop a flexibility strategy. In fact, any of the data-driven strategies developed throughout this thesis take these into account *implicitly*, and there are other simpler techniques which may be employed for variance reduction objectives[343]

8.4 Relative Performance Analysis

8.4.1 Distributional DRL Agents

The trends from the results confirm the trade-offs involved between average return and risk-tolerance which must be balanced by decision-makers. The distribution of performance on the same 2,000 testing scenarios for the agents trained as detailed in the former section is graphically captured by Figure 8.5. The results suggest distributional DRL to be highly effective at integrating decision-makers risk tolerance level when analysing flexibility in engineering systems design under uncertainty. As the level of risk sensitivity is increased, the respective CDF seems to become progressively narrower, with overall reduced variability in performance among scenarios. This is highly promising to increase the acceptability of the proposed data-driven methodology as compared to current approaches. Intuitively, the most risk-sensitive agents yield the worst overall expected performance across the tested environment episodes – consistent with the learning process depicted in Figure 8.3. More specifically, the ENPV was reduced by around 35% and 43% for the medium (α =

0.5) and highly ($\alpha = 0.25$) risk sensitive designs and flexibility strategies, respectively. In terms of upside potential, both of them can be visually observed to perform significantly worse than the risk-neutral agent. This is because the decision making policy is based on optimising the CVaR, thus does not consider those unusually high demand scenarios when developing operating strategies. Accordingly, however, the risk-sensitive designs substantially reduce downside exposure compared to the risk neutral one. The VaR improves by 13% and 21% for the medium and highly risk-sensitive designs. Both of them yielded very few instances of negative NPVs, which is likely to increase acceptability for risk-averse decision-makers or investors.

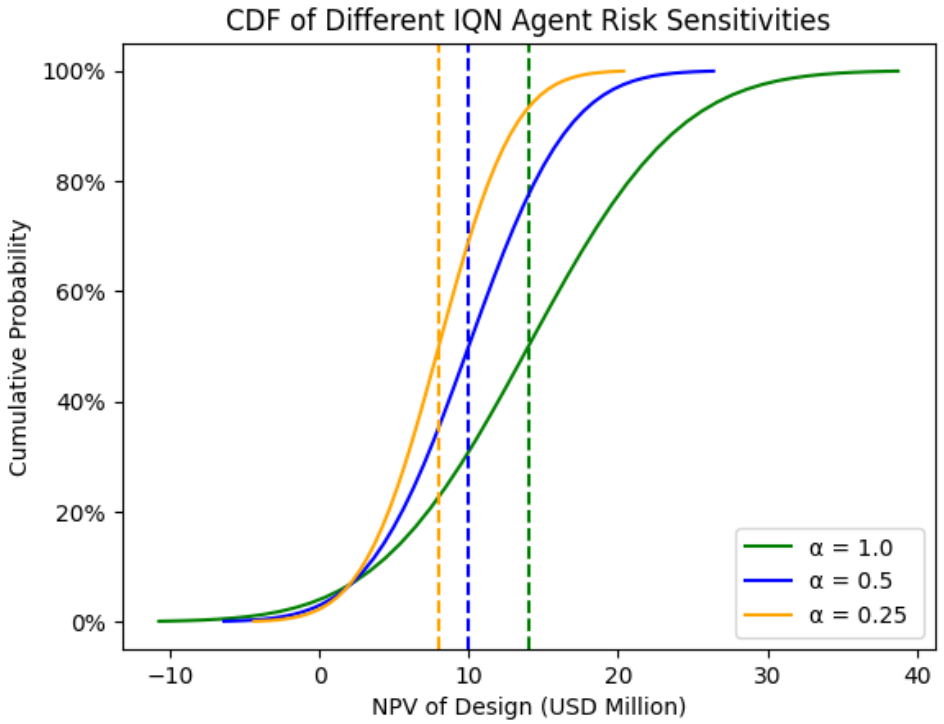


Figure 8.5: Cumulative distribution function showing expected performance of IQN agent with different risk-tolerance levels on the garage case study across 2000 simulated scenarios. Vertical dashed lines represent the ENPV for each agent.

8.4.2 Benchmarking against Standard Methods

Important improvements are also found compared to the other design solutions evaluated in Chapter 4 (with their results summarised by Table 4). The ENPV for the risk-neutral IQN implementation is very competitive (~2% lower) with the one obtained through DDQN in 4.4.1. This is notable given the non-distributional implementation used more than twice the training steps, even if wall-clock training time was overall similar between them. IQN agent performance can also likely be enhanced if augmented with any of the possible DQN variants. This suggest the value of approximating a return distribution even for risk-neutral objectives, and is consistent with the trends recorded in the original publication[339]. Furthermore, the VaR is substantially greater for the risk-sensitive agents presented here than for the standard DRL implementation, DP SAA, and Flexibility DRs - the key motivations

originally driving the work in this Chapter. Nonetheless, the ENPV of the risk neutral optimised designs via DP SAA and Flexibility DRs from 4.4.3 is greater than that obtained via the risk-sensitive IQN agents depicted in Figure 8.5. This is driven by the same trade-offs between average return and risk-tolerance which must be balanced by decision-makers. Finally, both risk-aware agents outperform the inflexible baseline on VaR and ENPV- albeit less focused on capturing upside potential.

8.5 Computational Insights for Risk-Tolerance Integration

8.5.1 Comparative Decision Rule Distribution

The policies developed by the IQN agents in this chapter can, as a starting point, be broken down into the retroactive decision rule distribution as was done 4.5.2. Note, however, that this does not give any direct insights into risk-tolerance integration, but rather serves to provide some initial intuition on how the improved performance at the tail end of the distributions is achieved here. The plots are not included again as they would be repetitive, albeit their key insights are briefly summarised below.

Firstly, the risk sensitive solutions are found to primarily deploy 1 or 2 levels in the zero stage decision (as opposed to 2 or 3 in the risk-neutral one). This leads to an average rule of $\boldsymbol{f_0 = 1.86}$ and $\boldsymbol{f_0 = 1.52}$ for the medium and highly risk-aware agents, respectively. Initial investments are thus substantially lower than those obtained via standard DRL Flexibility Analysis ($\boldsymbol{f_0 = 2.63}$), meaning the designs in this section are more focused on a modular deployment strategy than economies of scale. This already reduces project risk significantly, as capacity is truly deployed only if and when needed – discounting capital costs into the future even if missing out on some potential revenues in early years. Nonetheless, it should be noted that even more risk-sensitive DRL agents (not presented here) did not really reduce this any further, even if great uncertainty is recognized in the environment (see Figure 4.2). In fact, the demand model parameters used mean going much lower would intuitively be suboptimal decision-making, as there is nearly always sufficient demand to fill up the capacity provided by one or two floors even in starting years.

The number of years with demand greater than capacity before expansion also intuitively suggest the development of more risk-sensitive policies through a distributional perspective. Usually, demand exceeds capacity for two or three years before proceeding with expansion decisions. As a result, an average rule of $\boldsymbol{d_r = 1.97}$ and $\boldsymbol{d_r = 2.35}$ are recorded for the medium and highly risk-aware agents, respectively. These are slightly higher than the value obtained with standard DRL ($\boldsymbol{d_r = 1.76}$), as the risk-sensitive flexibility strategy further ensures any built capacity will be consistently filled to reduce downside exposure. A greater difference was initially expected in this regard, however this is likely correlated to the fact that lower capacities are built in the zero stage decision for the risk-sensitive agents. Accordingly, they rely more on modular deployment to capture uncertain demand realizations,

meaning there is likely to be more total expansion decisions within a single simulated scenario. Nonetheless, waiting longer before triggering system flexibility results in a reduced ability to capture upside potential, as there is more consistently several years with unmet demand (and thus potential revenues not collected).

These findings are further supported by the results for capacity adjustment magnitude on expansion years. Average rules of $\boldsymbol{f_t = 1.35}$ and $\boldsymbol{f_t = 1.21}$ are recorded for the medium and highly risk-aware agents, respectively. The reduction from the standard DRL strategy ($\boldsymbol{f_t = 1.67}$) is mainly driven by the risk-sensitive agents never expanding capacity by three floors to minimize investment risk – and as such less suited to capture revenues in rapidly rising demand profiles. Nonetheless, expansion by one floor was the most commonly selected action across all levels of risk-sensitivity considered.

The distribution of expansion timing is also aligned with the trends above and reduces the likelihood of overbuilding. In fact, flexibility strategies are very rarely triggered past year 15 for the risk-sensitive agents as it becomes increasingly difficult to consistently recoup capital investments made under uncertainty. This contrasts to the standard DRL agent which presented some probabilities of expansion even up to year 19, albeit reducing gradually past year 12, as shown in Figure 4.11.

8.5.2 Quantile Function Visualization

One of the most important contributions of a distributional perspective towards acceptability of the proposed approach is the improved visualization ability for decision-making. In fact, rather than simply providing an expected value for each action in each state, an entire return distribution becomes available. This helps give some intuition on the potential risk-level of each action, and refine DRL agent action selection to decision-makers risk-tolerance during system operation. The general logic is captured by Figure 8.6, showing estimated returns at different points in the project for the medium risk-sensitive agent (α = 0.5) only. The graphic points to the rich information gained by the distributional approach, where if looking at different quantiles different actions are optimal as the curves cross over each other. It also suggests how these distributions are adapted based on state observations, with substantially different shapes and suggestions output over time for the same scenario. In fact, if the risk-tolerance for an engineering system changes over time, these may be used to inform decision-makers of potential deviations from the trained policy to better suit their preferences. Furthermore, these could be visualized as they evolve across a single environment episode through a GIF format, improving the intuition behind these methods. This would be particularly useful in a design-decision support tool built around the proposed data-driven approach, for instance as introduced in 6.5.3.

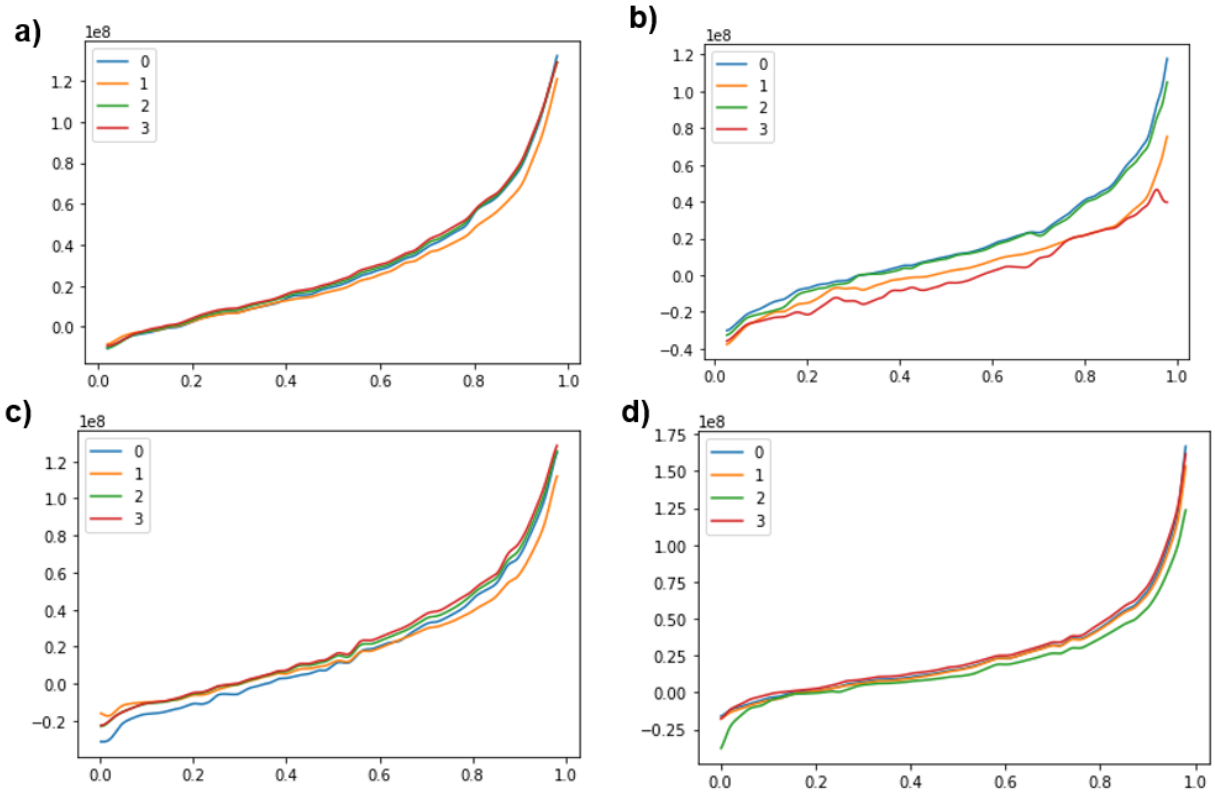


Figure 8.6: Example quantile functions at year 1 (a), year 20 (b), year 5 (c) and year 15 (d) showing the estimated returns (y-axis) for each quantile (x-axis) in a single stochastic scenario.

It is important to note in this context that, to maintain reasonable computational expense, only 6 samples (N and N' from Algorithm 5) were used to update the return distribution via quantile regression in each timestep. This was found to yield a good risk-sensitive performance while leading to a similar training time as the standard DRL method, as overall this is an engineering systems design and planning problem with a relatively low level of risk. In applications where the perceived risk level or uncertainty is higher (for instance the environments tackled in 5.3.1, 6.3.1, or 7.3.1), however, the number of samples could be increased to obtain a more granular representation of the return distribution. This choice could be adapted based on decision-makers preferences and project specific factors. In fact, increasing the computational budget in early design stages is ultimately a minor limitation in real-world projects - where this may result in differences in expected value in the range of millions of dollars, if not more.

8.6 Chapter Summary and Conclusions

This chapter focused on enhancing the acceptability of DRL Flexibility Analysis in practice by providing a dynamic and systematic approach to integrate decision-makers risk-profiles into the policy optimisation process. Distributional DRL, and in particular the IQN algorithm, was at the core of these results, where an entire return distribution is obtained for each action from each state observation as opposed to a single scalar value. The increased data allowed the formulation of a medium and highly risk-sensitive policy, showing very promising results in terms of VaR for the infrastructure design case study originally introduced in Chapter 4. As expected, some important trade-offs were recorded among risk-sensitivity and expected performance, which allow the tailoring of the proposed approach to different kinds of decision-makers. A comparative decision-rule distribution provided some valuable intuition on how strategies differed based on optimised risk-awareness. Quantile plot visualizations serve to further improve the acceptability of these methods and give a general blueprint for how they may be integrated into a decision-support tool in order to adaptively tailor decision-making to the perceived level of uncertainty.

There are several avenues for extending the work presented in this chapter. Firstly, application to higher dimensional case studies should be considered, keeping in mind that computational tractability was one of the original motivations for the DRL-based approach presented in this thesis. A broader spectrum of risk-tolerances could be evaluated, yielding a finer differentiation in decision-making among them. Importantly, ways to implement a distributional approach with policy-gradient based agents should be explored, as their advantages compared to value based ones discussed in earlier chapters still hold here. This could work very nicely by substituting the value function for the value return distribution function in the critic component of actor-critic algorithms, improving performance for both risk-neutral and risk-aware agents. Finally, while unable to be presented here due to time limitations, an alternative approach based on scenario clustering and training separate DRL agents for each estimated risk level (not using a distributional perspective) is also likely to be valuable. Such a method would aid decision-makers in better understanding the risk-level associated with a project and flexibility strategy, which can be very intuitive to non-domain practitioners. Nonetheless, this would be more focused on the infrastructure management point of view, as distributional DRL is deemed to provide more robust and data-driven avenue to deal with risk-sensitivity under uncertainty.

Chapter 9

9 Conclusions and Research Outlooks

9.1 Summary of Contributions

In conclusion, this thesis was successful in achieving its overall objective of developing a data-driven approach to analyse flexibility in engineering systems design under uncertainty. A generic framework for ways in which ML techniques may augment state of the art Flexibility in Design was proposed in Chapter 3, exploring a number of different applications. The greatest potential was recognized in the concept generation and design space exploration phases. Accordingly, Deep Reinforcement Learning Flexibility Analysis was developed, addressing many of the limitations of currently used approaches, while presenting important parallels with existing methods. The power of neural network function approximation in sequential decision making problems under many sources of uncertainty was found to be very valuable to the targeted objectives. More specifically, it allows for the development of flexibility strategies which recognize a much greater degree of uncertainty than alternative methods, as well as an overall deeper exploration of the design and decision-making space. Furthermore, this thesis showed a natural progression to progressively more complex application case studies, with Figure 9.1 highlighting the relevant computational expense and dimensionality considerations.

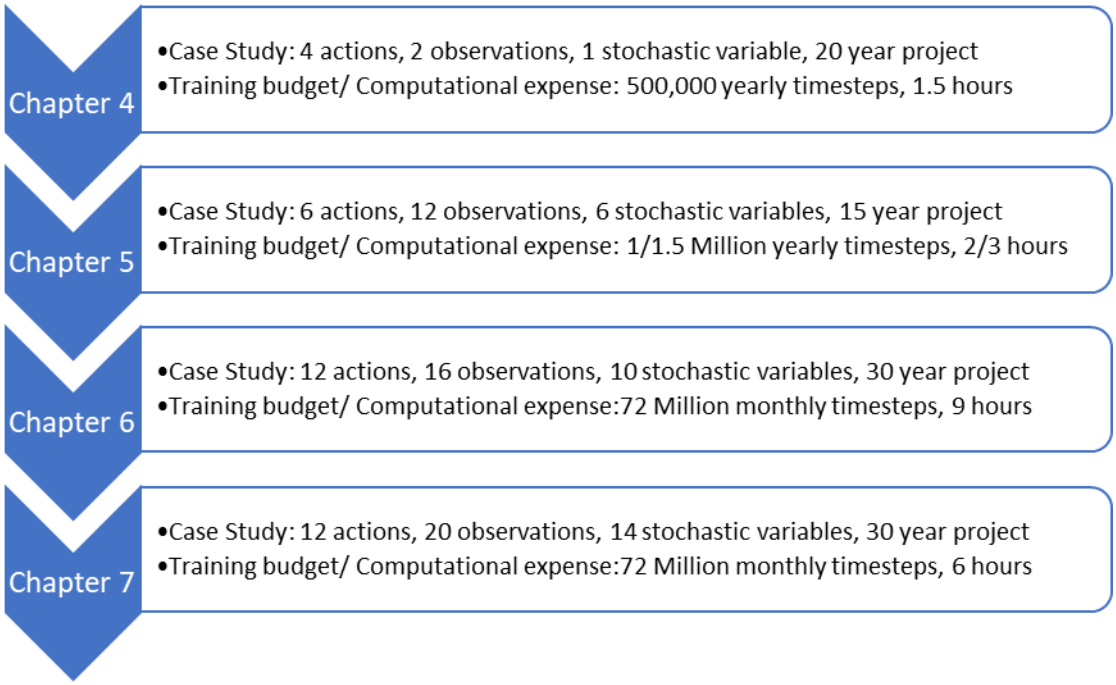


Figure 9.1: Summary of complexity and computational expense progression followed in this thesis.

The proposed approach was first validated on a simpler proof of concept case study of an infrastructure system which is used to teach the ideas around Flexibility in Design. The results presented in Chapter 4 demonstrated the potential of the proposed methodology, justifying an expansion to a more complex case study to further explore its potential contributions in Chapter 5. Here, a deeper analysis of the proposed approach was performed, looking at increasing decision-making flexibility from that considered with alternative methods. The initial objectives for the thesis were further met through the work in Chapter 6, as DRL Flexibility analysis allowed the first ever design and planning of a mobile microgrid targeted at global nomadic communities. This represented a case study whose dimensionality could not be tackled with standard methods, highlighting the contributions of the proposed approach to general engineering systems planning. In Chapter 7, a governmental policy-tool formulation perspective was taken (as opposed to system operators or designers in the previous chapters). This served to highlight how a data-driven perspective allows for the systematic consideration of uncertainty on market support instruments, and its impact on both optimal energy system configurations and creation of policy-tools based on targeted objectives. Finally, Chapter 8 presented a dynamic approach to integrate decision-makers risk profiles into the development of DRL based flexibility strategies, increasing their acceptability across various application contexts.

Based on the above, this research provides a significant contribution to the field of Flexibility in Design, and to general engineering systems planning under uncertainty. Several avenues for improving interpretability and acceptability are presented, providing an important contribution to the practical implementation of DRL models in general. The following subsections provide a more detailed overview of the most important conclusions from this thesis.

9.2 Trade-offs with Alternative Methods

The primary motivations for the work presented here were that a data-driven approach may allow a deeper exploration of the potential flexibility strategy space and the consideration of more sources of uncertainty and decision-making possibilities than with alternatives. The general DRL Flexibility Analysis methodology was defined for these objectives in Chapter 3, aiming to obtain better performing design solutions under uncertainty, and forming the primary focus of the thesis. The proposed approach was then validated on a proof-of-concept case study in Chapter 4, which is typically used to teach the ideas around Flexibility in Design, and has been extensively studied in the literature. Accordingly, the most extensive benchmarking with alternative methods was presented on this case study. The expected performance of DRL based solutions was compared to stochastically optimal inflexible baselines, DP, and flexibility decision-rules – representing the state-of-the art in the field. The results were highly promising, showing improved expected value under uncertainty was achieved through a more dynamic solution than possible with other approaches. As the thesis moved to methodological expansion (Chapter 5) and maturation (Chapter 6), the case studies tackled were

progressively more dimensional. As several sources of uncertainty and flexibility enablers were considered, benchmarking DRL solutions with standard methods became computationally intractable. This was a very important contribution as it allowed the optimisation of design problems more realistic to the actual encountered conditions. Given the lack of extensive benchmarking with other flexible solutions, however, the trade-offs necessarily had to be evaluated more qualitatively than in the validation phase. Nonetheless, design solutions developed through the proposed approach were found to outperform rigid baselines in all cases considered, highlighting the value of explicitly considering flexibility under the many sources of uncertainty impacting our society.

In terms of expected performance of design solutions under uncertainty, the results throughout this thesis suggest the proposed DRL based approach can consistently yield improvements from alternative methods. More specifically, when evaluating on the same subset of scenarios as optimised for, strategies can outperform those obtained with dynamic programming based methods or flexibility decision-rules. Solutions found were also shown to be more generalizable than with standard methods through several out-of-sample testing approaches, which can be attributed to the power of neural network function approximation. This is a crucial finding in light of the limited certainty of projections made in early design stages, encouraging engineering system designers in several fields to implement a data-driven perspective to planning and sizing. Furthermore, the computational expense using DRL was generally comparable (or even reduced) to alternatives, showing it is practical to apply it in real world conditions. This is particularly true when dealing with large capital investment projects, where even marginal improvements by a few percentage points can lead to increased value in the range of millions of dollars, if not more, depending on the system at hand. It should be noted, however, that as the decision-making flexibility was increased, so was the variability in performance across scenarios. This is intuitive as more options inherently make it more likely to select the wrong decision in highly stochastic environments. Nonetheless, the risk-profile integration approach presented in Chapter 8 addresses these concerns, and highlights the balance between average and worst case performance (i.e. VaR) which must be determined by engineering system designers.

The improvements in expected performance, however, sacrifice interpretability and acceptability to a certain degree. This is a finding that is true across all “black-box” optimisation techniques implemented in AI or ML, hence why they have not been yet applied in a lot of industrial settings. In the Flexibility in Design field, these trade-offs are particularly evident against flexibility decision-rules, which provide readily usable guidelines and triggering signals to exercise strategies during operation. Dynamic programming based methods, on the other hand, are well accepted in industry due to their global optimum convergence guarantees, but actually do not provide better interpretability than DRL based solutions. In fact, they share the same mathematical and theoretical foundation, but if uncertainty realizations deviate from the scenarios used during optimisation, the solutions do not generalize nearly as well as DRL Flexibility analysis ones. Furthermore, several avenues for

improving interpretability and acceptability were explored throughout this thesis, helping address some of these limitations. These were primarily based on advanced data-visualization to allow decision-makers to easily understand the logic behind DRL agent actions. Comparison to optimised predefined flexibility decision rules were used as a starting point due to the methodological similarities. Policy-gradient based methods then allowed visualization of action selection probabilities, and how actual state observations may be used during system operation to dynamically recommend strategies. Contributing to concept generation, the retroactive analysis of novel unintuitive strategies through advanced statistical decomposition of decision-making under different conditions was also shown. Furthermore, an initial prototype for a design-decision support system around the proposed approach was presented, which should make the findings more accessible to all sorts of system operators. Finally, the dynamic risk-profile integration potential shown in Chapter 8 marks a significant improvement to alternatives, helping increase acceptability in practice.

9.3 Summary of Limitations

As discussed in the former section, while there are some trade-offs to be considered with alternative methods, the results from this thesis suggest the proposed approach is superior under most metrics or evaluation criteria. One issue which has not yet been considered, however, is the simplicity of formulation within the proposed framework to non-domain practitioners. The results presented were based on a very extensive deep-dive into the field over 4 years, and most of them took many iterations to reach the performance presented here. In contrast, other Flexibility in Design approaches have been extensively studied and documented in the literature, thus providing an easier roadmap for non-domain practitioners to reproduce their performance. Creating the representative environment, for instance, was found to be significantly more complex than simply coding a mathematical model. This is because choices made during environment crafting can significantly affect training simplicity, learning stability, and final agent performance due to several important design choices.

The most important one is arguably reward metric formulation, as this is the primary feedback signal guiding DRL policy optimisation, akin to the objective function in alternative methods. All of the case studies presented in this thesis ultimately used economic based rewards, as they were deemed more representative of actual decision-makers objectives and financial realities. A simplistic approach of reward penalties were used to allow the DRL agents to implicitly learn the form of constraints. This worked well in the final implementations presented here, however took significant iterations to balance exploration vs exploitation of the decision-making space. Too large a penalty, in fact, was found to excessively hinder exploration and lead to significantly sub-optimal policies. A value too small, on the other hand, yielded agents more focused on exploitation, which would as a result tend to violate constraints more than acceptable in practice. Accordingly, no general conclusions could be drawn on this regard, and the final ones were found through empirical testing.

Similarly, the power of neural network function approximation introduced the consideration of several potential timestep resolutions for the same design problem. This different significantly from alternatives, where the number of decision-stages has to be limited to maintain computational tractability. Overall, this produced a trade-off between dynamicity of the developed policy and ease of training. For instance, monthly resolution was used for decision-making in the mobile microgrid case study from Chapter 6 and Chapter 7 to capture seasonal factors, while Chapters 4 and 5 implemented a yearly one. The impact of changing this is significant, as using a finer timestep resolution was found to generally improve solution performance as it can adapt engineering system configuration more dynamically to realized uncertainties. On the other hand, training becomes significantly more computationally intensive as the horizon increases, necessitating better differentiation between states to determine optimal action selecting.

The creation of the observation and action space presented many of the same complexities. In terms of observation, it is deemed important to only include variables which are representative and informative of engineering system condition. An excessive number of state features which are not really necessary for decision-makers was in practice found to lead to policies which assumed correlations which weren't there, as well as increasing the computational budget required for learning. On the contrary, omitting certain key variables from the state vector can lead to crucial interactions not being accounted for by the agent, and thus sub-optimal decision making over time and uncertainty. Similarly, the action space created should be representative of flexibility enablers to respond to stochastic realizations of key variables. Increasing dimensionality of the decision-making space corresponds to enhanced complexity in training, thus it ought to be done only in cases where additional options can improve expected value for the targeted engineering system. In environments where this is not necessarily intuitive a priori, therefore, the action space should be iteratively refined. This refinement can be based on empirical evaluations of training performance, starting with all decision-making possibilities, and progressively masking them to yield the appropriate balance between dynamic flexibility strategies and practical considerations.

There are also hundreds of potential DRL agent variants which can be used for developing policies under uncertainty, and the choice among them is not straightforward. In fact, state-of-the art results have been achieved through several different algorithms in the benchmark domain typically assessed in DRL literature. Generally, in the context of analysing flexibility for engineering systems, policy-gradient based methods were found to be superior to value-based ones. This was driven by their enhanced ability to handle high dimensional environments, as well as the possibility to develop stochastic policies which are better suited to decision-making under deep uncertainty. Nonetheless, approaches maintaining an explicit value function representation present several key advantages, namely training stability or interpretability in practice. Overall, therefore, algorithm selection introduces trade-offs between sample complexity, wall-clock training time, and expected performance

– among others. As such, there is no overarching optimal agent to be implemented, and it must be determined based on environment dimensionality, uncertainty recognition, and decision-makers risk tolerances. The approach taken in this thesis, which is suggested for future works implementing DRL Flexibility Analysis, is to gain some initial intuition on which class of methods is best suited to a particular problem by starting with open-source DRL agent libraries. These allow for fast prototyping and experimentation in the initial phases, helping to shortlist the best candidates. Following this, however, it is deemed important to actually code the agents from scratch in order to understand the logic behind them, and have more freedom on potential modifications to the original implementations.

Finally, reproducibility of results is highly sensitive to both neural network architectures and hyperparameters used for optimisation. Aiming to yield a more consistent and intuitive solution procedure, fully connected MLPs were implemented for all case studies presented in this thesis. While their depths and number of neurons in each layer were adapted based on problem space complexity, better performance was sometimes found using alternative architectures. In fact, the most famous DRL implementations used in industry at the moment use much more complex architectures, and explore different activation functions for different objectives. Furthermore, hyperparameter performance is highly correlated with the network architecture whose parameters they seek to optimise. While the Optuna package provided a straightforward approach to hyperparameter optimisation, the ranges considered still had to be manually set to maintain computational tractability. This introduces the potential for human bias or error, and some intuition on both environment nature and algorithms logic is essential to tune the potential hyperparameter space efficiently.

9.4 Direction for Future Research

As the work presented in this thesis has contributed to the development of a completely novel approach for the design and planning of engineering systems under uncertainty, several potentially valuable research outlooks have been introduced.

Firstly, an important motivation for a Flexibility in Design approach is that it may be a unifying paradigm for sustainability and resilience. These were explicitly evaluated in Chapters 6 and 7, confirming the validity of this paradigm in practice, but the right way to balance competing objectives could not be conclusively determined. In fact, the reward functions considered were purely financial across all case studies analysed in this thesis. Future work could look at multi-objective formulations to explicitly optimise all three objectives, and evaluate what additional gains may be made on sustainability and resilience metrics when directly integrated into the reward signals. This may be as simple as using a cost factor in the reward function to convert all metrics into monetary units, with their weights assigned based on decision-makers priorities. Additionally, separate neural network parametrized policies may be developed targeted at each objective, and implemented at different points during system operation based on realized uncertainties and system performance.

A library in Python could be developed specifically for DRL flexibility analysis, helping reduce some of the domain expertise requirements discussed in this section. This could for instance, automatically convert a mathematical model to an OpenAI gym environment to provide a starting point for analysis. This initial environment would likely still need to be iterated upon due to the importance of its formulation on the performance of the proposed approach, but could help facilitate experimentation with DRL to system designers not familiar with the field. The resulting improved simplicity would help make the novel method more accessible to engineering system designers and planners which currently employ standard methods. In turn, this would aid in recognizing its added value across a wide range of potential applications. Furthermore, such a package could be used to produce some initial recommendations for algorithm selection based on environment formulation, as well as any exogenous computational constraints. Similarly, it could allow for the streamlining of testing the performance of alternative and more complex neural network architectures, which may be better suited to approximate flexibility strategies for different problems. This could be extended to hyperparameter optimisation, providing a "one-stop-shop" to address the primary complexities associated with the proposed approach.

There are also several promising avenues for further enhancing the interpretability and acceptability of DRL Flexibility Analysis. The development of a generic graphical user interface (GUI) as part of a design-decision support system is of high priority. Such an approach would further extend the accessibility of the developed method even to users who are not familiar with Python or coding. The GUI could be used to visualize the logic behind implementing different flexibility strategies as uncertainty is resolved over time. This could be based on advanced statistical decomposition of implicit decision rules used, as well as more complex methods such as PCA to determine the most impactful variables and features in a data-driven manner. Furthermore, a GUI would allow for direct interactions with users to elicit their risk-tolerances and integrate them into the policy optimisation process more systematically. This is important as most users would not be able to simply quantify their risk-sensitivity between 0 and 1 as was done in Chapter 8. Different prompting techniques or decision-making scenarios may be used for this, although more advanced approaches should also be explored. Finally, if any assumptions or stochastic model parameters change during system operation, the GUI could be used to easily keep track of them over time. This could include, but not limited to, the development of novel market support instruments or uncertainties which necessarily change the optimality of the solution. Accordingly, the DRL agent policy could be easily updated based on these changing parameters, dynamically adapting decision-making based on realised conditions.

In summary, while there are several promising avenues for future work, this thesis provided a very important methodological contribution to analysing flexibility in engineering systems design under uncertainty. The majority of future work, therefore, would be focused on improving interpretability, accessibility, and acceptability to maximize its contributions in real-world system operation.

10 References

1 Blumrich, J. F. Design. *Science* **168** (1970).

2 Dieter, G. E. & Schmidt, L. C. *Engineering design*. (McGraw-Hill Higher Education Boston, 2009).

3 Marchau, V. A., Walker, W. E., Bloemen, P. J. & Popper, S. W. *Decision making under deep uncertainty: From theory to practice*. (Springer Nature, 2019).

4 Meng, D. *et al.* Multidisciplinary design optimization of engineering systems under uncertainty: a review. *International Journal of Structural Integrity* **13**, 565-593 (2022).

5 De Weck, O., Eckert, C. M. & Clarkson, P. J. in *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07. 2007.* 159-160 (exec. Summ.), full paper no. DS142_P_480.

6 Neufville, R. D., Smet, K., Cardin, M.-a. & Ranjbar-bourani, M. Decision Making under Deep Uncertainty. *Decision Making under Deep Uncertainty*, 223-252 (2019). https://doi.org/10.1007/978-3-030-05252-2

7 MacCormack, A., Herman, K. The Rise and Fall of Iridium. (Harvard Business School, Cambridge, MA, United States, 2001).

8 Lim, R. & Ng, K. in *The Business Times* (Marshall Cavendish Business Information, Singapore, 2011).

9 Shepard, W. An Update On China's Largest Ghost City - What Ordos Kangbashi Is Like Today. *Forbes* (2016). <http://www.forbes.com/sites/wadeshepard/2016/04/19/an-update-on-chinas-largest-ghost-city-what-ordos-kangbashi-is-like-today/#2a651eb91e08>.

10 Caputo, C. & Cardin, M.-A. Analyzing Real Options and Flexibility in Engineering Systems Design Using Decision Rules and Deep Reinforcement Learning. *Journal of Mechanical Design* **144**, 021705 (2021).

11 Caunhye, A. M. & Cardin, M.-A. An approach based on robust optimization and decision rules for analyzing real options in engineering systems design. *IISE Transactions* **49**, 753-767 (2017).
12 Trigeorgis, L. (Cambridge, MA: MIT Press, 1996).
13 Dixit, A. & Pindyck, R. in *Investment Under Uncertainty* 1-17 (Princeton University Press, 1994).
14 Suh, E. S., De Weck, O., Kim, I. Y. & Chang, D. Flexible platform component design under uncertainty. *Journal of Intelligent Manufacturing* **18**, 115-126 (2007). https://doi.org/10.1007/s10845-007-0008-x
15 Brundtland, G. H. Report of the World Commission on Environment and

Development: Our Common Future. (United Nations, 1987).
16 Roostaie, S., Nawari, N. & Kibert, C. Sustainability and resilience: A review of definitions, relationships, and their integration into a combined building assessment framework. *Building and Environment* **154**, 132-144 (2019).
17 Biermann, F., Kanie, N. & Kim, R. E. Global governance by goal-setting: the novel approach of the UN Sustainable Development Goals. *Current Opinion in Environmental Sustainability* **26**, 26-31 (2017).
18 Breitkopf, P. & Coelho, R. F. *Multidisciplinary design optimization in computational mechanics*. (John Wiley & Sons, 2013).
19 Righi, A. W., Saurin, T. A. & Wachs, P. A systematic literature review of resilience engineering: Research areas and a research agenda proposal. *Reliability Engineering & System Safety* **141**, 142-152 (2015).
20 Bhamra, R., Dani, S. & Burnard, K. Resilience: the concept, a literature review and future directions. *International journal of production research* **49**, 5375-5393 (2011). https://doi.org/10.1080/00207543.2011.563826
21 Rodriguez-Nikl, T. Linking disaster resilience and sustainability. *Civil Engineering and Environmental Systems* **32**, 157-169 (2015).
22 Achour, N., Pantzartzis, E., Pascale, F. & Price, A. D. F. Integration of resilience and sustainability: from theory to application. https://doi.org/10.1108/IJDRBE-05-2013-0016
23 Turner Ii, B. Vulnerability and resilience: Coalescing or paralleling approaches for sustainability science? *Global Environmental Change* **20**, 570-576 (2010).
24 Redman, C. L. Should sustainability and resilience be combined or remain distinct pursuits? *Ecology and Society* **19** (2014).
25 Larsen, L. *et al.* (USGBC).
26 Ahern, J. From fail-safe to safe-to-fail: Sustainability and resilience in the new urban world. *Landscape and urban Planning* **100**, 341-343 (2011).
27 Bocchini, P., Frangopol, D. M., Ummenhofer, T. & Zinke, T. Resilience and Sustainability of Civil Infrastructure: Toward a Unified Approach. *Journal of Infrastructure Systems* **20** (2014). https://doi.org/10.1061/(ASCE)IS.1943-555X.0000177
28 Waage, S. A. Re-considering product design: a practical "road-map" for integration of sustainability issues. *Journal of Cleaner production* **15**, 638-649 (2007).
29 Rose, A. Resilience and sustainability in the face of disasters. *Environmental Innovation and Societal Transitions* **1**, 96-100 (2011).
30 Padgett, J. E. & Tapia, C. Sustainability of natural hazard risk mitigation: Life cycle analysis of environmental indicators for bridge infrastructure. *Journal of Infrastructure Systems* **19**, 395-408 (2013).
31 Lee, M. & Basu, D. An Integrated Approach for Resilience and Sustainability in Geotechnical Engineering. *Indian Geotechnical Journal* **48**, 207-234 (2018). https://doi.org/10.1007/s40098-018-0297-3
32 Comber, M. V., Poland, C. & Sinclair, M. in *Structures Congress 2012* (eds John Carrato & Joseph Burns) 910-921 (Reston, VA: American Society of Civil Engineers, Reston, VA, 2012).
33 Tapia, C., Ghosh, J. & Padgett, J. E. in *Structures Congress 2011.* 1937-1948.

34 Rodriguez-Nikl, T., Christiansen, J. W. & Walters, K. Reliability-based life cycle assessment of green concrete structures. *Special Publication* **289**, 1-13 (2012).
35 Berkes, F. & Folke, C. in *Workshop Property rights and the performance of natural Resource systems.*
36 Rosa, L. V., França, J. E., Haddad, A. N. & Carvalho, P. V. A Resilience Engineering Approach for Sustainable Safety in Green Construction. *Journal of Sustainable Development of Energy, Water and Environment Systems* **5**, 480-495 (2017).
37 Wied, M., Oehmen, J. & Welo, T. Conceptualizing resilience in engineering systems: An analysis of the literature. *Systems Engineering* (2019).
38 Chester, M. V. & Allenby, B. Toward adaptive infrastructure: flexibility and agility in a non-stationarity age. *Sustainable and Resilient Infrastructure* **4**, 173-191 (2019).
39 Hertzler, G. & Harris, M. in *Institute for Environmental Decisions Conference: Environmental Decisions: Risks and Uncertainties.*
40 Bosch-Badia, M.-T., Montllor-Serrats, J. & Tarrazon-Rodon, M.-A. Corporate social responsibility: A real options approach to the challenge of financial sustainability. *PloS one* **10**, e0125972 (2015).
41 Kim, S., Giachetti, R. & Park, S. Real Options Analysis for Acquisition of New Technology: A Case Study of Korea K2 Tank's Powerpack. *Sustainability* **10**, 3866 (2018).
42 de Neufville, R. & Smet, K. in *Decision Making under Deep Uncertainty* 117-132 (Springer International Publishing, 2019).
43 Trigeorgis, L. *Real options: Managerial flexibility and strategy in resource allocation*. (MIT press, 1996).
44 Ahmad, A. M. *The use of refurbishment, flexibility, standardisation and BIM to support the design of a change-ready healthcare facility*, Loughborough University, (2014).
45 Maseda, L. J. *Real option analysis of flexibility in a hospital emergency department expansion project: a systems approach*, Massachusetts Institute of Technology, (2008).
46 Asokan, V. A., Yarime, M. & Esteban, M. Introducing flexibility to complex, resilient socio-ecological systems: A comparative analysis of economics, flexible manufacturing systems, evolutionary biology, and supply chain management. *Sustainability (Switzerland)* **9**, 1-17 (2017). https://doi.org/10.3390/su9071091
47 de Neufville, R., Lee, Y. S. & Scholtes, S. in *IEEE Conference on Infrastructure Systems.* 8-10 (Rotterdam, Netherlands).
48 De Neufville, R. & Scholtes, S. *Flexibility in engineering design*. (MIT Press, 2011).
49 Cardin, M.-A. *et al.* Empirical evaluation of procedures to generate flexibility in engineering systems and improve lifecycle performance. *Research in Engineering Design* **24**, 277-295 (2013).
50 Pierce, J. G. Designing Flexible Engineering Systems Utilizing Embedded Architecture options. *Dissertation*, 16 (2010).
51 Martínez-Ceseña, E. A. & Mutale, J. Application of an advanced real options approach for renewable energy generation projects planning. *Renewable and Sustainable Energy Reviews* **15**, 2087-2094 (2011). https://doi.org/10.1016/j.rser.2011.01.016
52 de Weck, O., de Neufville, R. & Chaize, M. 23-25.
53 Martínez-Costa, C., Mas-Machuca, M., Benedito, E. & Corominas, A. A review of mathematical programming models for strategic capacity planning in manufacturing. *International Journal of Production Economics* **153**, 66-85 (2014).
54 Lehtonen, M. & Nye, S. History of electricity network control and distributed generation in the UK and Western Denmark. *Energy Policy* **37**, 2338-2345 (2009).
55 Van der Vleuten, E. & Lagendijk, V. Interpreting transnational infrastructure vulnerability: European blackout and the historical dynamics of transnational electricity governance. *Energy Policy* **38**, 2053-2062 (2010).
56 World Bank, I. (DataBank, 2020).
57 ESMAP, W. B. Mini-Grids-for-Half-a-Billion-People-Market-Outlook-and-Handbook-for-Decision-Makers. (2019).

58 Tenenbaum, B., Greacen, C., Siyambalapitiya, T. & Knuckles, J. *From the bottom up: how small power producers and mini-grids can deliver electrification and renewable energy in Africa*. (The World Bank, 2014).
59 Bhattacharyya, S. C. & Palit, D. *Mini-grids for rural electrification of developing countries: analysis and case studies from South Asia*. (Springer, 2014).
60 Yadoo, A. & Cruickshank, H. The role for low carbon electrification technologies in poverty reduction and climate change strategies: A focus on renewable energy mini-grids with case studies in Nepal, Peru and Kenya. *Energy Policy* **42**, 591-602 (2012).
61 Cardin, M.-A. Enabling flexibility in engineering systems: a taxonomy of procedures and a design framework. *Journal of Mechanical Design* **136** (2014).
62 Cardin, M.-A., Xie, Q., Ng, T. S., Wang, S. & Hu, J. An approach for analyzing and managing flexibility in engineering systems design based on decision rules and multistage stochastic programming. *IISE Transactions* **49**, 1-12 (2017).
63 Cardin, M.-A., Zhang, S. & Nuttall, W. J. Strategic Real Option and Flexibility Analysis for Nuclear Power Plants Considering Uncertainty in Electricity Demand and Public Acceptance. *Energy Economics* **64**, 226-237 (2017). https://doi.org/10.1016/j.eneco.2017.03.023
64 Zhang, S. & Cardin, M.-A. Flexibility and real options analysis in emergency medical services systems using decision rules and multi-stage stochastic programming. *Transportation Research Part E: Logistics and Transportation Review* **107**, 120--140 (2017). https://doi.org/10.1016/j.tre.2017.09.003
65 Jonathan Mak, W. H. & John Clarkson, P. in *Procedia CIRP* Vol. 60 536-541 (Elsevier B.V., 2017).
66 Rolnick, D. *et al.* Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433* (2019).
67 Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction*. (MIT press, 2018).
68 Li, Y. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274* (2017).
69 Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* **34**, 26-38 (2017).
70 Evans, R. & Gao, J. Deepmind ai reduces google data centre cooling bill by 40%. *DeepMind blog* **20**, 158 (2016).
71 Manyika, J. An overview of Bard: an early experiment with generative AI. (Technical report, Google AI, 2023).
72 OpenAI. Vol. 2023 (ed OpenAI) (OpenAI, Online, 2023).
73 Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
74 Liu, L., Zhang, M. & Zhao, Z. The Application of Real Option to Renewable Energy Investment: A Review. *Energy Procedia* **158**, 3494-3499 (2019). https://doi.org/10.1016/j.egypro.2019.01.921
75 Hino, M. & Hall, J. W. Real options analysis of adaptation to changing flood risk: Structural and nonstructural measures. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* **3**, 04017005 (2017).
76 Karan, E., Asadi, S., Mohtar, R. & Baawain, M. Towards the optimization of sustainable food-energy-water systems: A stochastic approach. *Journal of cleaner production* **171**, 662-674 (2018).
77 Gao, J. & You, F. Design a sustainable supply chain under uncertainty using life cycle optimisation and stochastic programming. *CHEMICAL ENGINEERING* **61** (2017).
78 Namany, S., Al-Ansari, T. & Govindan, R. Sustainable energy, water and food nexus systems: A focused review of decision-making tools for efficient resource management and governance. *Journal of Cleaner Production* **225**, 610-626 (2019).
79 Lee, K. & Kum, D. Complete design space exploration of isolated hybrid renewable energy system via dynamic programming. *Energy conversion and management* **196**, 920-934 (2019).
80 Hu, J., Guo, P. & Poh, K. L. Flexible capacity planning for engineering systems based on decision rules and differential evolution. *Computers & Industrial Engineering* **123**, 254-262 (2018).

81 Ding, H., Pinson, P., Hu, Z. & Song, Y. Optimal offering and operating strategies for wind-storage systems with linear decision rules. *IEEE Transactions on Power Systems* **31**, 4755-4764 (2016).
82 Dominguez, R., Conejo, A. J. & Carrion, M. Investing in generation capacity: A multi-stage linear-decision-rule approach. *IEEE Transactions on Power Systems* **31**, 4784-4794 (2016).
83 Zhang, J., Wang, Z. & Zhang, H. Data-based optimal control of multiagent systems: A reinforcement learning design approach. *IEEE transactions on cybernetics* **49**, 4441-4449 (2018).
84 Camburn, B., He, Y., Raviselvam, S., Luo, J. & Wood, K. Machine learning-based design concept evaluation. *Journal of Mechanical Design* **142** (2020).
85 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206-215 (2019).
86 Perera, A. T. D., Wickramasinghe, P., Nik, V. M. & Scartezzini, J.-L. Introducing reinforcement learning to the energy system design process. *Applied Energy* **262**, 114580 (2020).
87 Lee, X. Y., Balu, A., Stoecklein, D., Ganapathysubramanian, B. & Sarkar, S. A case study of deep reinforcement learning for engineering design: Application to microfluidic devices for flow sculpting. *Journal of Mechanical Design, Transactions of the ASME* **141**, <xocs:firstpage xmlns:xocs=""/> (2019). https://doi.org/10.1115/1.4044397
88 Wang, T. & De Neufville, R. in *real options conference, Paris, France.* (Citeseer).
89 Bowman, E. H. & Moskowitz, G. T. in *Source: Organization Science.* 772-777.
90 Nilchiani, R. & Hastings, D. E. Measuring the Value of Flexibility in Space Systems: A Six-Element Framework. *Systems Engineering* **10**, 26-44 (2007).
91 Mikaelian, T., Nightingale, D. J., Rhodes, D. H. & Hastings, D. E. Real Options in Enterprise Architecture: A Holistic Mapping of Mechanisms and Types for Uncertainty Management. *IEEE Transactions on Engineering Management* **54**, 457-470 (2011).
92 Pahl, G. & Beitz, W. *Engineering design: a systematic approach*. (Springer Science & Business Media, 2013).
93 Suh, N. P. & Suh, P. N. *The principles of design*. (Oxford University Press on Demand, 1990).
94 Hatchuel, A. & Weil, B. CK design theory: an advanced formulation. *Research in engineering design* **19**, 181-192 (2009).
95 Kurtoglu, T., Tumer, I. Y. & Jensen, D. C. A functional failure reasoning methodology for evaluation of conceptual system architectures. *Research in Engineering Design* **21**, 209-234 (2010).
96 Howard, T. J., Culley, S. & Dekoninck, E. A. Reuse of ideas and concepts for creative stimuli in engineering design. *Journal of Engineering Design* **22**, 565-581 (2011).
97 Howard, T., Culley, S. J. & Dekoninck, E. in *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07. 2007.* 329-330 (exec. Summ.), full paper no. DS342_P_493.
98 Thompson, G. & Lordan, M. A review of creativity principles applied to engineering design. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering* **213**, 17-31 (1999).
99 Charyton, C., Jagacinski, R. J., Merrill, J. A., Clifton, W. & DeDios, S. Assessing creativity specific to engineering with the revised creative engineering design assessment. *Journal of Engineering Education* **100**, 778-799 (2011).
100 Mikaelian, T. *An integrated real options framework for model-based identification and valuation of options under uncertainty*, Massachusetts Institute of Technology, (2009).
101 Gil, N. On the value of project safeguards: Embedding real options in complex products and systems. *Research Policy* **36**, 980-999 (2007).
102 Giffin, M. *et al.* Change propagation analysis in complex technical systems. *Journal of Mechanical Design* **131** (2009).
103 Bartolomei, J. E., Hastings, D. E., De Neufville, R. & Rhodes, D. H. Engineering Systems Multiple-Domain Matrix: An organizing framework for modeling large-scale complex systems. *Systems Engineering* **15**, 41-61 (2012).

104 Hu, J. & Cardin, M.-A. Generating flexibility in the design of engineering systems to enable better sustainability and lifecycle performance. *Research in Engineering Design* **26**, 121-143 (2015).
105 Chechurin, L. & Collan, M. *Advances in Systematic Creativity*. (Springer, 2019).
106 Shah, N. B., Viscito, L., Wilds, J., Ross, A. M. & Hastings, D. E. in *6th Conference on Systems Engineering Research, Los Angeles, CA.* (Citeseer).
107 Caputo, C. & Cardin, M.-A. The role of machine learning for flexibility and real options analysis in engineering systems design. *Proceedings of the Design Society* **1**, 3121-3130 (2021).
108 Nemirovski, A., Juditsky, A., Lan, G. & Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* **19**, 1574-1609 (2009).
109 Prékopa, A. *Stochastic programming*. Vol. 324 (Springer Science & Business Media, 2013).
110 Wang, T. & de Neufville, R. in *INCOSE International Symposium.* 1124-1133 (Wiley Online Library).
111 Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P. & Bertsekas, D. P. *Dynamic programming and optimal control*. Vol. 1 (Athena scientific Belmont, MA, 1995).
112 Bartolomei, J. E., De Neufville, R., Hastings, D. E. & Rhodes, D. H.
113 Cardin, M.-A. & Geltner, D.
114 Garstka, S. J. & Wets, R. J.-B. On decision rules in stochastic programming. *Mathematical Programming* **7**, 117-143 (1974).
115 Arciszewski, T. Engineering semantic evaluation of decision rules. *Journal of Intelligent & Fuzzy Systems* **5**, 285-295 (1997).
116 Alpaydin, E. *Introduction to machine learning*. (MIT press, 2020).
117 Mosavi, A. *et al.* State of the art of machine learning models in energy systems, a systematic review. *Energies* **12**, 1301 (2019).
118 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436-444 (2015).
119 Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (MIT press, 2016).
120 Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks* **61**, 85-117 (2015).
121 Deng, L. & Yu, D. Deep learning: methods and applications. *Foundations and trends in signal processing* **7**, 197-387 (2014).
122 Amari, S.-i. Backpropagation and stochastic gradient descent method. *Neurocomputing* **5**, 185-196 (1993).
123 Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* **12** (2011).
124 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
125 Hinton, G., Srivastava, N. & Swersky, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on* **14**, 2 (2012).
126 Panchal, J. H., Fuge, M., Liu, Y., Missoum, S. & Tucker, C. Machine Learning for Engineering Design. *Journal of Mechanical Design* **141** (2019).
127 Vázquez-Canteli, J. R. & Nagy, Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy* **235**, 1072-1089 (2019).
128 Shahan, D. & Seepersad, C. C. in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.* 303-313.
129 Sharpe, C., Wiest, T., Wang, P. & Seepersad, C. C. A comparative evaluation of supervised machine learning classification techniques for engineering design applications. *Journal of Mechanical Design* **141** (2019).
130 Wang, P., Wang, Z. & Almaktoom, A. T. Dynamic reliability-based robust design optimization with time-variant probabilistic constraints. *Engineering Optimization* **46**, 784-809 (2014).
131 Lee, J. H., Shin, J. & Realff, M. J. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering* **114**, 111-121 (2018).

132 Toivonen, H. & Gross, O. Data mining and machine learning in computational creativity. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **5**, 265-275 (2015).
133 Markov, A. A. The theory of algorithms. *Trudy Matematicheskogo Instituta Imeni VA Steklova* **42**, 3-375 (1954).
134 Caputo, C. *et al.* Design and planning of flexible mobile Micro-Grids using Deep Reinforcement Learning. *Applied Energy* **335**, 120707 (2023).
135 Silver, D. *et al.* Mastering the game of go without human knowledge. *nature* **550**, 354-359 (2017).
136 Watkins, C. J. & Dayan, P. Q-learning. *Machine learning* **8**, 279-292 (1992).
137 Jaakkola, T., Jordan, M. & Singh, S. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems* **6** (1993).
138 Mnih, V. *et al.* Human-level control through deep reinforcement learning. *nature* **518**, 529-533 (2015).
139 Van Hasselt, H., Guez, A. & Silver, D. in *Proceedings of the AAAI Conference on Artificial Intelligence.*
140 Wang, Z. *et al.* in *International conference on machine learning.* 1995-2003 (PMLR).
141 Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. *arXiv preprint arXiv:1511.05952* (2015).
142 Bellemare, M. G., Dabney, W. & Munos, R. in *International Conference on Machine Learning.* 449-458 (PMLR).
143 Hessel, M. *et al.* in *Proceedings of the AAAI conference on artificial intelligence.*
144 Ghavamzadeh, M., Mannor, S., Pineau, J. & Tamar, A. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436* (2016).
145 Sutton, R. S., McAllester, D., Singh, S. & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* **12** (1999).
146 Andrew, N. (Youtube, Coursera, 2017).
147 Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **8**, 229-256 (1992).
148 Levine, S. & Koltun, V. in *International conference on machine learning.* 1-9 (PMLR).
149 Schulman, J., Levine, S., Abbeel, P., Jordan, M. & Moritz, P. in *International conference on machine learning.* 1889-1897 (PMLR).
150 Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438* (2015).
151 Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
152 Mnih, V. *et al.* in *International conference on machine learning.* 1928-1937 (PMLR).
153 Salimans, T., Ho, J., Chen, X., Sidor, S. & Sutskever, I. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864* (2017).
154 Lillicrap, T. P. *et al.* Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
155 Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. in *International conference on machine learning.* 1861-1870 (PMLR).
156 Kumar, H., Koppel, A. & Ribeiro, A. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, 1-35 (2023).
157 keras (https://github.com/fchollet/keras, 2015).
158 Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow. org (2015). *URL https://www. tensorflow. org* (2015).
159 Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).
160 Bradbury, J. *et al.* JAX: composable transformations of Python+ NumPy programs. (2018).
161 Brockman, G. *et al.* Openai gym. *arXiv preprint arXiv:1606.01540* (2016).
162 Duan, Y., Chen, X., Houthooft, R., Schulman, J. & Abbeel, P. in *International conference on machine learning.* 1329-1338 (PMLR).
163 Dhariwal, P. *et al.* (2017).

164 Guadarrama, S. *et al.* TF-Agents: A library for reinforcement learning in tensorflow. *see https://github. com/tensorflow/agents* (2018).
165 Castro, P. S., Moitra, S., Gelada, C., Kumar, S. & Bellemare, M. G. Dopamine: A research framework for deep reinforcement learning. *arXiv preprint arXiv:1812.06110* (2018).
166 Contributors, S. B. (https://stable-baselines.readthedocs.io/en/master/, 2020).
167 Liaw, R. *et al.* Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* (2018).
168 Bergstra, J., Yamins, D. & Cox, D. D. in *Proceedings of the 12th Python in science conference.* 20 (Citeseer).
169 (2020).
170 Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.* 2623-2631.
171 Haj-Ali, A. *et al.* A View on Deep Reinforcement Learning in System Optimization. (2019).
172 Mowbray, M., Petsagkourakis, P., del Rio-Chanona, E. A. & Zhang, D. Safe chance constrained reinforcement learning for batch process control. *Computers & chemical engineering* **157**, 107630 (2022).
173 Mowbray, M. *et al.* Machine learning for biochemical engineering: A review. *Biochemical Engineering Journal* **172**, 108054 (2021).
174 Petsagkourakis, P., Sandoval, I. O., Bradford, E., Zhang, D. & del Rio-Chanona, E. A. Reinforcement learning for batch bioprocess optimization. *Computers & Chemical Engineering* **133**, 106649 (2020).
175 Zhang, Z., Zhang, D. & Qiu, R. C. Deep reinforcement learning for power system applications: An overview. *CSEE Journal of Power and Energy Systems* **6**, 213-225 (2019).
176 Yonekura, K. & Hattori, H. Framework for design optimization using deep reinforcement learning. *Structural and Multidisciplinary Optimization* **60**, 1709-1713 (2019).
177 Cui, H., Turan, O. & Sayer, P. Learning-based ship design optimization approach. *Computer-Aided Design* **44**, 186-195 (2012).
178 Baker, B., Gupta, O., Naik, N. & Raskar, R. Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167* (2016).
179 Van Moffaert, K., Drugan, M. M. & Nowé, A. in *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL).* 191-199 (IEEE).
180 Qi, X., Wu, G., Boriboonsomsin, K., Barth, M. J. & Gonder, J. Data-driven reinforcement learning–based real-time energy management system for plug-in hybrid electric vehicles. *Transportation Research Record* **2572**, 1-8 (2016).
181 Kuznetsova, E. *et al.* Reinforcement learning for microgrid energy management. *Energy* **59**, 133-146 (2013).
182 O'Neill, D., Levorato, M., Goldsmith, A. & Mitra, U. in *2010 First IEEE international conference on smart grid communications.* 409-414 (IEEE).
183 Cardin, M.-A. & Caputo, C. in *Supergen Energy Networks Hub Risk Day.* (ed EPSRC) (Supergen).
184 De Neufville, R., Scholtes, S. & Wang, T. Real options by spreadsheet: parking garage case example. *Journal of infrastructure systems* **12**, 107-111 (2006).
185 Sarkar, S. *et al.* Multifidelity and Multiscale Bayesian Framework for High-Dimensional Engineering Design and Calibration. *Journal of Mechanical Design* **141** (2019). https://doi.org/10.1115/1.4044598
186 Shu, D. *et al.* 3D Design Using Generative Adversarial Networks and Physics-Based Validation. *Journal of Mechanical Design* **142** (2019). https://doi.org/10.1115/1.4045419
187 Market, C. *Multistory Parking Garage*, <https://www.pinterest.com/pin/790381803336997669/> (2020).
188 Largue, P. Go ahead for $1.5bn waste-to-energy plant at Singapore's flagship environmental facility. *Power Engineering International* (2020). <https://www.powerengineeringint.com/renewables/green-light-for-1-5bn-waste-to-energy-facility-at-singapores-tuas-nexus-iwmf1/>.

189 Wu, Y. N., Arto Emerik; Salminen,Pekka Kalevi; Johansen,Peter; Cui,Rong. Paving the Way to a Sustainable Heating Sector : A Roadmap for Ulaanbaatar Urban Heating (English). (World Bank, online, 2019).
190 NovaTerra. NovaTerra Response Datasheet. (Submitted to Imperial College London, 2020).
191 Tugjamba, N., Walkerden, G. & Miller, F. Adaptation strategies of nomadic herders in northeast Mongolia: climate, globalisation and traditional knowledge. *Local Environment* **26**, 411-430 (2021). https://doi.org/10.1080/13549839.2021.1891032
192 *Air pollution in Mongolia: Opportunities for further actions | United Nations Development Programme*.
193 Dai, X., Xu, Y., Jia, G. & Fu, C. in *IOP Conference Series. Earth and Environmental Science.* (IOP Publishing).
194 Group, W. B. (ed World Bank Group) (2021).
195 Otsuki, T. Costs and benefits of large-scale deployment of wind turbines and solar PV in Mongolia for international power exports. *Renewable Energy* **108**, 321-335 (2017).
196 Tong, J. C.-K., Tse, J. M.-Y. & Jones, P. J. Development of thermal evaluation tool for detached houses in Mongolia. *Energy and Buildings* **173**, 81-90 (2018).
197 Wilkin, S. *et al.* Economic Diversification Supported the Growth of Mongolia's Nomadic Empires. *Scientific Reports* **10**, 3916 (2020). https://doi.org/10.1038/s41598-020-60194-0
198 Yang, X. *et al.* Effects of climate change on cultivation patterns and climate suitability of spring maize in Inner Mongolia. *Sustainability* **13**, 8072 (2021).
199 GCRF, NovaTerra & Energy, M. o. in *Energy for Development – low-carbon energy and industry for economic growth in Mongolia Workshop 1* (Global Challenges Research Fund).
200 Chimed, G. Energy Sector of Mongolia, Country Report. *Energy Sector of Mongolia, Country Report* (2019).
201 Purev, U.-U. & Hagishima, A. A Field Survey of Traditional Nomadic Dwelling Gers Used as Urban Habitats in Ulaanbaatar, Mongolia. *Evergreen* **7**, 155-171 (2020).
202 Sovacool, B. K., D'Agostino, A. L. & Bambawale, M. J. Gers gone wired: lessons from the renewable energy and rural electricity access project (REAP) in Mongolia. *Energy for Sustainable Development* **15**, 32-40 (2011).
203 Mocellin, J. & Foggin, P. Health status and geographic mobility among semi-nomadic pastoralists in Mongolia. *Health & Place* **14**, 228-242 (2008). https://doi.org/10.1016/j.healthplace.2007.06.005
204 Tamir, K., Urmee, T. & Pryor, T. Issues of small scale renewable energy systems installed in rural Soum centres in Mongolia. *Energy for Sustainable Development* **27**, 1-9 (2015).
205 (IEA), I. E. A. *Mongolia Electricity Market Report*, <https://www.iea.org/countries/mongolia> (2021).
206 Bank, A. D. Mongolia: Upscaling Renewable Energy Sector Project. (INTEGRATION environment & energy GmbH, Ulaanbaatar, Mongolia, 2018).
207 Park, H., Fan, P., John, R., Ouyang, Z. & Chen, J. Spatiotemporal changes of informal settlements: Ger districts in Ulaanbaatar, Mongolia. *Landscape and Urban Planning* **191**, 103630 (2019). https://doi.org/10.1016/j.landurbplan.2019.103630
208 Cardin, M.-A. & Hu, J. Analyzing the tradeoffs between economies of scale, time-value of money, and flexibility in design under uncertainty: Study of centralized versus decentralized waste-to-energy systems. *Journal of Mechanical Design* **138** (2016).
209 Caunhye, A. M. & Cardin, M.-A. in *6th International Symposium on Reliability Engineering and Risk Management* (Singapore, 2018).
210 de Neufville, R., Smet, K., Cardin, M.-A. & Ranjbar-Bourani, M. in *Decision making under deep uncertainty* 223-252 (Springer, Cham, 2019).
211 Gallagher, C. V., Bruton, K., Leahy, K. & O'Sullivan, D. T. The suitability of machine learning to minimise uncertainty in the measurement and verification of energy savings. *Energy and Buildings* **158**, 647-655 (2018).
212 Mavromatidis, G., Orehounig, K. & Carmeliet, J. Design of distributed energy systems under uncertainty: A two-stage stochastic programming approach. *Applied energy* **222**, 932-950 (2018).

213 Fioriti, D., Poli, D., Duenas-Martinez, P. & Perez-Arriaga, I. Multi-year stochastic planning of off-grid microgrids subject to significant load growth uncertainty: overcoming single-year methodologies. *Electric Power Systems Research* **194**, 107053 (2021).
214 Postali, F. A. & Picchetti, P. Geometric Brownian motion and structural breaks in oil prices: a quantitative analysis. *Energy Economics* **28**, 506-522 (2006).
215 Dufresne, D. The integral of geometric Brownian motion. *Advances in Applied Probability* **33**, 223-241 (2001).
216 Marathe, R. R. & Ryan, S. M. On the validity of the geometric Brownian motion assumption. *The Engineering Economist* **50**, 159-192 (2005).
217 PhD Thesis Source Code (GitHub, 2023).
218 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261-272 (2020).
219 Bagaram, M. B. & Tóth, S. F. Multistage sample average approximation for harvest scheduling under climate uncertainty. *Forests* **11**, 1230 (2020).
220 Pardo, F., Tavakoli, A., Levdik, V. & Kormushev, P. in *International Conference on Machine Learning.* 4045-4054.
221 Petsagkourakis, P., Sandoval, I. O., Bradford, E., Zhang, D. & Chanona, E. A. d. R. Constrained reinforcement learning for dynamic optimization under uncertainty. *arXiv preprint arXiv:2006.02750* (2020).
222 Zhang, C., Vinyals, O., Munos, R. & Bengio, S. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893* (2018).
223 Blume, S. O. P., Sansavini, G. & Cardin, M.-A. Fuzzy control-enabled flexible short-turning for real-time disruption management in urban transit systems. *Manuscript in Preparation* (2021).
224 EPA. Scope 1 and Scope 2 Inventory Guidance. (Greenhouse Gas Protocol, WASHINGTON, D.C, USA, 2020).
225 Ge, P., Teng, F., Konstantinou, C. & Hu, S. A Resilience-Oriented Centralised-to-Decentralised Framework for Networked Microgrids Management. *arXiv preprint arXiv:2109.00245* (2021).
226 Pogaku, N., Prodanovic, M. & Green, T. C. Modeling, analysis and testing of autonomous operation of an inverter-based microgrid. *IEEE Transactions on power electronics* **22**, 613-625 (2007).
227 Ge, P., Zhu, Y., Green, T. C. & Teng, F. Resilient secondary voltage control of islanded microgrids: An eskbf-based distributed fast terminal sliding mode control approach. *IEEE Transactions on Power Systems* **36**, 1059-1070 (2020).
228 Riverso, S., Sarzo, F. & Ferrari-Trecate, G. Plug-and-play voltage and frequency control of islanded microgrids with meshed topology. *IEEE Transactions on Smart Grid* **6**, 1176-1184 (2014).
229 Nasir, M., Khan, H. A., Hussain, A., Mateen, L. & Zaffar, N. A. Solar PV-based scalable DC microgrid for rural electrification in developing regions. *IEEE Transactions on sustainable energy* **9**, 390-399 (2017).
230 Che, L., Zhang, X., Shahidehpour, M., Alabdulwahab, A. & Abusorrah, A. Optimal interconnection planning of community microgrids with renewable energy sources. *IEEE Transactions on Smart Grid* **8**, 1054-1063 (2015).
231 Elsayed, A. T., Mohamed, A. A. & Mohammed, O. A. DC microgrids and distribution systems: An overview. *Electric power systems research* **119**, 407-417 (2015).
232 Mohammed, A., Refaat, S. S., Bayhan, S. & Abu-Rub, H. Ac microgrid control and management strategies: Evaluation and review. *IEEE Power Electronics Magazine* **6**, 18-31 (2019).
233 Jadav, K. A., Karkar, H. M. & Trivedi, I. A Review of microgrid architectures and control strategy. *Journal of The Institution of Engineers (India): Series B* **98**, 591-598 (2017).
234 Borazjani, P., Wahab, N. I. A., Hizam, H. B. & Soh, A. B. C. in *2014 IEEE Innovative Smart Grid Technologies-Asia (ISGT ASIA).* 749-753 (IEEE).
235 Pfenninger, S. & Staffell, I. *Renewables Ninja*, <https://www.renewables.ninja/about> (2019).

236 Pfenninger, S. & Staffell, I. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. *Energy* **114**, 1251-1265 (2016).
237 Staffell, I. & Pfenninger, S. Using bias-corrected reanalysis to simulate current and future wind power output. *Energy* **114**, 1224-1239 (2016).
238 Gelaro, R. *et al.* The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of climate* **30**, 5419-5454 (2017).
239 Co., S. J. W. C. (Alibaba, 2021).
240 Company, C. B. E. A. (Alibaba, 2021).
241 Yueqing Jyins Electric Technology Co., L. (Alibaba, Alibaba, 2021).
242 Alibaba. (2019).
243 Ltd., R. S. E. C. (Alibaba, 2021).
244 Zhejiang Zhiyi Power Supply Technology Co., L. (Alibaba, Alibaba, 2021).
245 Co., S. F. T. (Alibaba, Alibaba, 2021).
246 Equipment, Q. G. N. E. (Alibaba, 2021).
247 Coalition, C. a. C. A. Financing Househehold Clean Heating Solutions. (UN Environment Finance Initiative, 2018).
248 ESMAP. *Energy Sector Management Assistance Program (ESMAP) Annual Report* (Washington, D.C., 2021).
249 Gu, W. *et al.* Modeling, planning and optimal energy management of combined cooling, heating and power microgrid: A review. *International Journal of Electrical Power & Energy Systems* **54**, 26-37 (2014).
250 IRENA. Renewable Power Generation Costs in 2020. (International Renewable Energy Agency, Abu Dhabi, 2021).
251 Borchers, A. M., Duke, J. M. & Parsons, G. R. in *Energy Policy* Vol. 35 (2007).
252 Kakinuma, K., Yanagawa, A., Sasaki, T., Rao, M. P. & Kanae, S. Socio-ecological interactions in a changing climate: a review of the Mongolian pastoral system. *Sustainability* **11**, 5883 (2019).
253 Byambadorj, T., Amati, M. & Ruming, K. J. Twenty-first century nomadic city: Ger districts and barriers to the implementation of the Ulaanbaatar City Master Plan. *Asia Pacific Viewpoint* **52**, 165-177 (2011).
254 Fernandez-Gimenez, M. E., Batjav, B. & Baival, B. Lessons from the Dzud. (2012).
255 Acosta, M. N. *et al.* Optimal under-frequency load shedding setting at Altai-Uliastai regional power system, Mongolia. *Energies* **13**, 5390 (2020).
256 Song, Y. Disclosable Version of the ISR - Ulaanbaatar Clean Air Project - P122320 - Sequence No : 19 (English). (World Bank Group, Washington, D.C, 2021).
257 Allen, R. W. *et al.* An assessment of air pollution and its attributable mortality in Ulaanbaatar, Mongolia. *Air Quality, Atmosphere & Health* **6**, 137-150 (2013).
258 Running, I., Sellers, D., Ramaraj, P. & McMullin, P. in *2020 IEEE Conference on Technologies for Sustainability (SusTech).* 1-5 (IEEE).
259 Maddalena, R. L. *et al.* Quantifying Space Heating Stove Emissions Related to Different Use Patterns in Mongolia. *Energy and Environment Research* **4**, 147 (2014).
260 Cowlin, S., Kaufmann, R. B., Edwards, R. & Smith, K. R. Impact of improved stoves on indoor air quality in Ulaanbaatar, Mongolia. (2005).
261 Bayandelger, B.-E., Ueda, Y. & Adiyabat, A. Experimental investigation and energy performance simulation of Mongolian Ger with ETS heater and solar PV in Ulaanbaatar city. *Energies* **13**, 5840 (2020).
262 Dashkhuu, D., Kim, J. P., Chun, J. A. & Lee, W.-S. Long-term trends in daily temperature extremes over Mongolia. *Weather and Climate Extremes* **8**, 26-33 (2015).
263 World Bank Group, E., Technical University of Denmark and Vortex. (2020).
264 World Bank Group, E., and SolarGIS. (2020).
265 Achiam, J., Held, D., Tamar, A. & Abbeel, P. in *International conference on machine learning.* 22-31 (PMLR).
266 Bhatnagar, S., Ghavamzadeh, M., Lee, M. & Sutton, R. S. Incremental natural actor-critic algorithms. *Advances in neural information processing systems* **20**, 105-112 (2007).

267 Munos, R., Stepleton, T., Harutyunyan, A. & Bellemare, M. G. Safe and efficient off-policy reinforcement learning. *arXiv preprint arXiv:1606.02647* (2016).
268 Wang, Z. *et al.* Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224* (2016).
269 Kuremoto, T., Hirata, T., Obayashi, M., Mabu, S. & Kobayashi, K. in *Time Series Analysis* (IntechOpen, 2019).
270 Andriotis, C. & Papakonstantinou, K. Managing engineering systems with large state and action spaces through deep reinforcement learning. *Reliability Engineering & System Safety* **191**, 106483 (2019).
271 Tomin, N., Zhukov, A. & Domyshev, A. in *EPJ Web of Conferences* Vol. 217 01016 (EDP Sciences, 2019).
272 Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research* **18**, 629-681 (2017).
273 Bhandari, J. & Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786* (2019).
274 DeVore, R., Hanin, B. & Petrova, G. Neural network approximation. *Acta Numerica* **30**, 327-444 (2021).
275 Bartlett, P. L., Montanari, A. & Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica* **30**, 87-201 (2021).
276 Henderson, P., Romoff, J. & Pineau, J. Where did my optimum go?: An empirical analysis of gradient descent optimization in policy gradient methods. *arXiv preprint arXiv:1810.02525* (2018).
277 Liu, Y., Gao, Y. & Yin, W. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems* **33**, 18261-18271 (2020).
278 Slater, H., De Boer, D., Qian, G. & Wang, S. in *Beijing: China Carbon Forum. http://www. chinacarbon. info/sdm_downloads/2020-china-carbon-pricing-survey.*
279 Pietzcker, R. C., Osorio, S. & Rodrigues, R. Tightening EU ETS targets in line with the European Green Deal: Impacts on the decarbonization of the EU power sector. *Applied Energy* **293**, 116914 (2021).
280 Lecocq, F. & Ambrosi, P. The Clean Development Mechanism: history, status, and prospects. *Review of Environmental Economics and Policy* **1**, 134-151 (2007).
281 Alex, G. *DE4 Master's Final Report –Designing an Intuitive Decision-Support System for the Operation of Future Energy Systems* MEng thesis, Imperial College London, (2022).
282 Blank, J. & Deb, K. Pymoo: Multi-objective optimization in python. *IEEE Access* **8**, 89497-89509 (2020).
283 in *Rome and the Worlds beyond its Frontiers* (eds Daniëlle Slootjes & M. Peachin) 96-122 (BRILL, 2019).
284 Westreicher, C. A., Mérega, J. L. & Palmili, G. The Economics of Pastoralism: Study on Current Practices in South America. *Nomadic Peoples* **11**, 87-105 (2007). https://doi.org/10.3167/np.2007.110205
285 Tsovoodavaa, G. & Kistelegdi, I. Comparative analysis for traditional yurts using thermal dynamic simulations in Mongolian climate. *Pollack Periodica* **14**, 97-108 (2019).
286 Cesare Caputo , P. G., Michel-Alexandre Cardin, Fei Teng, Anna Korre, Antonio del Rio Chanona. Enabling Nomadic Energy Access via Plug and Play Control and Deep Reinforcement Learning. *Nature Energy* (2023).
287 Sachs, J., Kroll, C., Lafortune, G., Fuller, G. & Woelm, F. *Sustainable development report 2022*. (Cambridge University Press, 2022).
288 IEA.
289 Ferris, E. Research on climate change and migration where are we and where are we going? *Migration Studies* **8**, 612-625 (2020). https://doi.org/10.1093/migration/mnaa028
290 Johnson, D. L. The nature of nomadism. *The nature of nomadism.* (1969).
291 Piguet, E., Pécoud, A. & De Guchteneire, P. Migration and climate change: An overview. *Refugee Survey Quarterly* **30**, 1-23 (2011).

292 Hoffmann, R., Dimitrova, A., Muttarak, R., Crespo Cuaresma, J. & Peisker, J. A meta-analysis of country-level studies on environmental change and migration. *Nature Climate Change* **10**, 904-912 (2020).
293 HOUSE, T. W. (ed White House) (Online, 2021).
294 Subbiah, A., Mansoor, S., Misra, R., Jaffer, H. & Tiwary, R. Addressing developmental needs through energy access in informal settlements. *Field Actions Science Reports. The journal of field actions*, 80-91 (2016).
295 Sayar, F. State of the Global Mini-grids Market Report 2020. *J Chem Inform Model* **53**, 1689-1699 (2019).
296 Boait, P. J. in *Mini-Grids for Rural Electrification of Developing Countries* 37-61 (Springer, 2014).
297 Mandelli, S., Barbieri, J., Mereu, R. & Colombo, E. Off-grid systems for rural electrification in developing countries: Definitions, classification and a comprehensive literature review. *Renewable and Sustainable Energy Reviews* **58**, 1621-1646 (2016).
298 Ge, P., Caputo, C., Cardin, M.-A., Korre, A. & Teng, F. A Wireless-Assisted Hierarchical Framework to Accommodate Mobile Energy Resources. *arXiv preprint arXiv:2209.09353* (2022).
299 Ge, P., Chen, B. & Teng, F. Event-triggered distributed model predictive control for resilient voltage control of an islanded microgrid. *International Journal of Robust and Nonlinear Control* **31**, 1979-2000 (2021).
300 Wang, Y., Rousis, A. O. & Strbac, G. On microgrids and resilience: A comprehensive review on modeling and operational strategies. *Renewable and Sustainable Energy Reviews* **134**, 110313 (2020).
301 Lian, J., Zhang, Y., Ma, C., Yang, Y. & Chaima, E. A review on recent sizing methodologies of hybrid renewable energy systems. *Energy Conversion and Management* **199**, 112027 (2019).
302 Siddaiah, R. & Saini, R. A review on planning, configurations, modeling and optimization techniques of hybrid renewable energy systems for off grid applications. *Renewable and Sustainable Energy Reviews* **58**, 376-396 (2016).
303 United Nations Department of Economic and Social Affairs, S. D. SDG Indicators. (UN, Online, 2022).
304 Upton, C. Living off the land: Nature and nomadism in Mongolia. *Geoforum* **41**, 865-874 (2010).
305 Roeckert, J. & Kraehnert, K. Extreme Weather Events and Internal Migration: Evidence from Mongolia. *Economics of Disasters and Climate Change* **6**, 95-128 (2022). https://doi.org/10.1007/s41885-021-00100-8
306 Hahn, A. Complexity of Mongolian stakeholders' dzud preparation and response. *Natural Hazards* **92**, 127-143 (2018). https://doi.org/10.1007/s11069-017-2907-2
307 *Environment & air pollution*.
308 *Ulaanbaatar Air Quality Improvement Program: Report and Recommendation of the President*, 2018).
309 Foundation, T. A. (Ulaanbaatar, Mongolia: The Asia Foundation, 2014).
310 Barquet, K., Trimmer, C., Sturesson, A., Joyce, B. & Jambal, D. Piloting the SDG Synergies approach in Mongolia. *Stockholm, Sweden: Stockholm Environment Institute* (2019).
311 He, X., Wang, R., Wu, J. & Li, W. Nature of power electronics and integration of power conversion with communication for talkative power. *Nature Communications* **11**, 2479 (2020). https://doi.org/10.1038/s41467-020-16262-0
312 (!!! INVALID CITATION !!! 18-20,52).
313 Cheng, I. H., Johansen, P. & La Rocca, R. Mongolia: Development impacts of solar-powered electricity services. (The World Bank, 2014).
314 ESMAP. *BEYOND CONNECTIONS*

Energy Access Redefined through a Multi-Tier Framework Tracking Progress Towards Sustainable Energy Goals, 2015).

315 Rosato, A. *et al.* A Solar Thermal Application for Mongolian Detached Houses: An Energy, Environmental, and Economic Analysis Based on Dynamic Simulations. *Buildings* **9**, 185 (2019).
316 Chen, P. *et al.* The heterogeneous role of energy policies in the energy transition of Asia–Pacific emerging economies. *Nature Energy* **7**, 588-596 (2022).
317 Wang, Q. & Chen, Y. Barriers and opportunities of using the clean development mechanism to advance renewable energy development in China. *Renewable and Sustainable Energy Reviews* **14**, 1989-1998 (2010).
318 Batsukh, E. *et al.* Green Financing in Developing Countries: Experiences from Mongolia, Kenya and Nigeria. *Emerging Issues on Trade and Sustainability* **37** (2019).
319 van der Ploeg, F. Carbon pricing under uncertainty. *International Tax and Public Finance* **28**, 1122-1142 (2021).
320 Engstrom, L. *et al.* Implementation matters in deep policy gradients: A case study on ppo and trpo. *arXiv preprint arXiv:2005.12729* (2020).
321 Ma, J. Entropy Augmented Reinforcement Learning. *arXiv preprint arXiv:2208.09322* (2022).
322 Cesare Caputo , O. R. G., Michel-Alexandre Cardin, Bryan T Adey. Risk-sensitive management of infrastructure systems under uncertainty by combinining scenario clustering with Deep Reinforcement Learning. *Journal of Infrastructure Asset Management* (2023).
323 Sharma, N., Tabandeh, A. & Gardoni, P. Resilience analysis: A mathematical formulation to model resilience of engineering systems. *Sustainable and Resilient Infrastructure* **3**, 49-67 (2018).
324 Yodo, N. & Wang, P. Engineering Resilience Quantification and System Design Implications: A Literature Survey. *Journal of Mechanical Design* **138** (2016). https://doi.org/10.1115/1.4034223
325 Park, J., Seager, T. P., Rao, P. S. C., Convertino, M. & Linkov, I. Integrating risk and resilience approaches to catastrophe management in engineering systems. *Risk Analysis* **33**, 356-367 (2013).
326 Francis, R. & Bekera, B. A metric and frameworks for resilience analysis of engineered and infrastructure systems. *Reliability engineering & system safety* **121**, 90-103 (2014). https://doi.org/10.1016/j.ress.2013.07.004
327 Righi, A. W., Saurin, T. A. & Wachs, P. in *Reliability Engineering and System Safety* Vol. 141 142-152 (Elsevier Ltd, 2015).
328 Cao, Y., Tan, Y., Li, C. & Rehtanz, C. Chance-constrained optimization-based unbalanced optimal power flow for radial distribution networks. *IEEE transactions on power delivery* **28**, 1855-1864 (2013).
329 MacKenzie, C. A. & Zobel, C. W. Allocating resources to enhance resilience, with application to superstorm sandy and an electric utility. *Risk Analysis* **36**, 847-862 (2016).
330 Fang, Y.-P. & Sansavini, G. Optimum post-disruption restoration under uncertainty for enhancing critical infrastructure resilience. *Reliability Engineering & System Safety* **185**, 1-11 (2019).
331 Bistline, J. E. & Weyant, J. P. Electric sector investments under technological and policy-related uncertainties: a stochastic programming approach. *Climatic change* **121**, 143-160 (2013).
332 Giahi, R., MacKenzie, C. A. & Hu, C. Design optimization for resilience for risk-averse firms. *Computers & Industrial Engineering* **139**, 106122 (2020).
333 Bellemare, M. G., Dabney, W. & Rowland, M. *Distributional reinforcement learning*. (MIT Press, 2023).
334 Lyle, C., Bellemare, M. G. & Castro, P. S. in *Proceedings of the AAAI Conference on Artificial Intelligence.* 4504-4511.
335 Bernhard, J., Pollok, S. & Knoll, A. in *2019 IEEE Intelligent Vehicles Symposium (IV).* 2148-2155 (IEEE).
336 Théate, T. & Ernst, D. Risk-sensitive policy with distributional reinforcement learning. *Algorithms* **16**, 325 (2023).

337 Liu, C., van Kampen, E.-J. & De Croon, G. C. in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 7198-7204 (IEEE).
338 Dabney, W., Rowland, M., Bellemare, M. & Munos, R. in *Proceedings of the AAAI Conference on Artificial Intelligence*.
339 Dabney, W., Ostrovski, G., Silver, D. & Munos, R. in *International conference on machine learning*. 1096-1105 (PMLR).
340 Wang, S. S. A class of distortion operators for pricing financial and insurance risks. *Journal of risk and insurance*, 15-36 (2000).
341 Chow, Y. & Ghavamzadeh, M. Algorithms for CVaR optimization in MDPs. *Advances in neural information processing systems* **27** (2014).
342 Hoel, C.-J., Wolff, K. & Laine, L. Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* (2023).
343 Greensmith, E., Bartlett, P. L. & Baxter, J. Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. *Journal of Machine Learning Research* **5** (2004).
344 National Environmental Agency. *Solid waste management*, <http://app2.nea.gov.sg/energy-waste/waste-management/refuse-disposal-facility> (2013).
345 Shell. *Shell Station Price Board*, <http://www.shell.com.sg/products-services/on-the-road/fuels/price-board.html> (2014).
346 National Environment Agency. Environmental Protection Division Report (2011).
347 IUT Global Pte Ltd. 9.5 MW food waste based grid connected power project implemented by IUT Singpaore Pte Ltd. (Clean Development Mechanism, 2006).
348 RIS international Ltd. Feasibility of Generating Green Power through Anaerobic Digestion of Garden Refuse from the Sacramento Area. (MacViro Consultants, 2005).
349 Singapore Power. *Electricity fariff rate*, <http://www.singaporepower.com.sg/irj/portal?NavigationTarget=navurl://41c8e6a3faf48bb168af2c222faa8ee4&windowId=WID1366188757420> (2013).
350 Bai, R. & Sutanto, M. The practice and challenges of solid waste management in Singapore. *Waste Management* **22**, 557-567 (2002).
351 Evangelisti, S., Lettieri, P., Borello, D. & Clift, R. Life cycle assessment of energy from waste via anaerobic digestion: A UK case study. *Waste Management* **34**, 226-237 (2014).

11 Appendix

11.1 Appendix I: Model assumptions for infrastructure chapter 4 case study

Model Parameters and Assumptions for Garage case study from Chapter 4

λ	12%
σ_{demand}	15%
D_0	750/spaces/year
D_f	1750/spaces/year
p	USD 1000 /space

c_r	USD 2000/space/year
c_l	USD 3600000/year
c_0	USD 16000/space
g_p	10% / year
f_{max}	9 floors
α	1000
β	0.15

11.2 Appendix II: Hyperparameters used for DDQN agent training in chapter 4.

DDQN Agent Hyperparameters Optimized via Optuna (not stated in Chapter 4)

η	0.000003	Buffer size	100,000
λ	0.99	Exploration fraction	0.7
Exploration initial ϵ	1.0	Exploration final ϵ	0.1
Batch size	32	Learning starts	10000
τ	500	α_{PE}	0.6
ϵ_{PE}	0.35	Memory sampling	Every 4 steps

11.3 Appendix III: Model assumptions for WTE case study from chapter 5

Table 14: Assumptions and parameters used in WTE model. All values are taken from the study by Cardin and Hu for benchmarking purposes [40].

Parameters	Definition	Assumptions	Comments and Source
C_{dis}	unit cost for disposing residues	S$77 /ton	This is the dispose cost which should be paid by AD plants to dispose of residues in incineration plants [344].
Cap_v	vehicle capacity for collecting wastes per trip	25 tonnes	
C_{fuel}	unit cost for fuel consumption	S$0.4/km	The price of diesel fuel is S$ 1.625 in Singapore [345]. It is assumed one liter of diesel fuel can last for 4km of travel distance.
D_{coi}	distance for collecting wastes within sector i	54km	It is assumed that the distances for collecting waste within the 6 sectors are

			the same. This distance is assumed based on Google map
D_{Tri}	distance for transporting wastes from sector i to the main sector	0km, 20km, 25km, 29km, 36km, 40km	0 km represents the transporting distance of the main site. It is assumed that no additional effort is required to transporting the wastes to the main site if the wastes are collected within the main site. The rest numbers are represented as the distances from other sectors to the main site and are assumed based on Google map.
$d^t(d_i^t)$	amount of recycled food wastes at year t (in sector i)	274 tpd (ton per day)	The total recycled food wastes in Singapore in 2013 is 100,000 tonnes [346].
E_g	electricity generation rate	230 kwh/ton	It is assumed that the biogas generation rate is 150 m^3/t [347]. The electricity conversion rate is 35% and only 80% of the generated electricity can be sold to power grid [348].
P_{to}	unit tipping fee for food wastes	S$65 /ton	It is assumed to be slightly lower than C_{dis} to encourage organic waste separation
P_e	unit selling price for electricity	S$0.27/kwh	It is estimated based on the Singapore electricity tariff in 2013 [349].
K	Coefficient parameter for cost function	305,288	It is estimated based on the real data from [348]. Detailed analysis can be found in [104]. A reasonable range of economies of scale factor is 0.6 to 1.0.
α	economies of scale factor	0.8	
ρ	unit land rental fee for the installed capacity	S$816/tpd	It is generated based on [62].
Σ	residues rate for food wastes	5%	The residues rate for incineration technology is 10% in Singapore [350]. The resides rate for AD is assumed to be less than 10% since it has high efficiency [351].
Ω	purity rate for food wastes	70%	In Singapore, the food wastes from industrial and commercial areas have about 30%-40% impurities [8]
T	lifecycle period	15	Long-term lifecycle period
λ	discount rate	8%	A general discounted rate

11.4 Appendix IV: Hyperparameters used for TRPO agent training in chapter 5

TRPO Agent Hyperparameters Optimized via Optuna (not stated in Chapter 5)

η	0.000003	Timesteps per batch	1024

λ	0.99	D_{KL}^{max}	0.1
Lam	0.97	Conjugate gradient iterations	20
Entropy coefficient	0.03	Conjugate gradient damping factor	0.023
Value function step size	0.000027	Value function iterations	5
ϵ_{PE}	0.35	Memory sampling	Every 4 steps

11.5 Appendix V: Detailed mathematical formulation of power flow model from chapter 6 (Note that equation numbering is separate from main text)

$$P_{NEL}{}^{n,t} = ED^{n,t} - P_{res}^{n,t} \quad (1)$$

$$P_{res}^{n,t} = P_{wind,adj}^{n,t} + P_{pv,adj}^{n,t} \quad (2)$$

$$P_{Load,Net}{}^{n,t} = HD^{n,t} + ED^{n,t} - P_{res}^{n,t} \quad (3)$$

$$P_{HD,eh,res}{}^{n,t} = \min\left(HD^{n,t}, P_{eh,max}{}^{n,t}, -P_{ED,Load,Net}{}^{n,t}\right) \quad , P_{NEL}{}^{n,t} \leq 0 \quad (4)$$

$$P_{NHL}^{n,t} = HD^{n,t} - P_{eh,res}{}^{n,t} \quad (5)$$

$$P_{res,PL}^{n,t} = P_{res}^{n,t} - P_{HD,eh,res}{}^{n,t} - P_{NEL}{}^{n,t} \quad , P_{ED,Load,Net}{}^{n,t}, P_{HD,Load,Net}^{n,t} \leq 0 \quad (6)$$

$$n \in [N_1, , N_2, N_3, .. N_N], \forall n, t \quad (7)$$

$$P_{ba,\ out\ max}^{n,t} = \theta_{ba}^{n,t} C_{out} \quad (8)$$

$$P_{ba,in,max}^{n,t} = \theta_{ba}^{n,t} C_{in} \quad (9)$$

$$P_{ba}^{n,t} = \begin{cases} -\min\left(P_{ba,out,max}^{n,t}, P_{NEL}{}^{n,t}\right) & P_{NEL}{}^{n,t} \geq 0 \quad (\boldsymbol{a}) \\ -\min\left(P_{ba,out,max}^{n,t}, P_{HD,Load,Net}^{n,t}\right) & P_{res,PL}^{n,t}, P_{ED,Load,Net}{}^{n,t} \leq 0 \; ; \; P_{HD,Load,Net}^{n,t} \geq 0 \; (\mathbf{b}) \\ \min\left(P_{ba,in,max}^{n,t}, P_{res,PL}^{n,t}\right) & P_{res,PL}^{n,t} \geq 0 \; ; \; P_{HD,Load,Net}^{n,t}, P_{ED,Load,Net}{}^{n,t} \leq 0 \quad (\boldsymbol{c}) \end{cases} \quad 10$$

$$E_{ba}^{n,t} = E_{ba}^{n,t-1} + P_{ba}^{n,t}\sqrt{\eta_{ba}} + E_{ba,new}^{n,t} \quad 11$$

$$E_{ba}^{n,t} = E_{ba}^{n,t-1} + P_{ba}^{n,t}\sqrt{\eta_{ba}} + E_{ba,new}^{n,t} \quad 12$$

$$P_{ul,PB}^{n,t} = \begin{cases} P_{Load,Net}{}^{n,t} - \left(E_{ba}^{n,t} - E_{ba,min}^{n,t}\right), & P_{Load,Net}{}^{n,t} \geq 0 \\ 0 & otherwise \end{cases} \quad 13$$

$$P_{HD,Load,Net,PB}^{n,t} = HD^{n,t} - P_{eh,res}{}^{n,t} - P_{eh,batt}{}^{n,t} \quad 14$$

$$P_{HD,Load,Net,PB,ES}^{t} = \sum_{n \in N} P_{HD,Load,Net,PB}^{n,t} \quad 15$$

$$P_{eg}^{n,t} = \min\left(P_{res,PL}^{n,t} - P_{ba}^{n,t}, 0\right) \qquad P_{ba}^{n,t} \geq 0 \quad 16$$

$$n \in N_{deficit} : P_{eg}^{n,t} \leq 0 \quad 17$$

$$n \in N_{surplus} : P_{eg}^{n,t} > 0 \quad 18$$

$$P_{ul,es}^{t} = \sum_{n \in N_{deficit}} P_{ul,PB}^{n,t} + P_{HD,Load,Net,PB}^{n,t} \quad 19$$

$$P_{uel,es,bpp}^{t} = \sum_{n \in N_{deficit}} P_{ul,PB}^{n,t} \quad 20$$

$$P_{eg,es}^{t} = \sum_{n \in N_{surplus}} P_{eg}^{n,t} \quad 21$$

$$P_{dist,es}^{t} = \min\left(P_{ul.es}^{t}, P_{eg,es}^{t}, \theta_{PPmg}^{t}, P_{dist,es,max}^{t}\right) \quad 22$$

$$s.t.\ 0\ \leq P_{dist,es}^{t} \leq 0.3\ (ED_{es}^{t}) \quad (23)$$

$$L_{dist}{}^{t,n} = \frac{L_{cb}^{t}}{N} \quad (24)$$

$$P_{dist}^{t,n} = \frac{P_{dist,es}^{t}}{N} \quad (25)$$

$$R_{dist,ref}^{t,n} = \frac{\rho_{cb} L_{dist}{}^{t,n}}{A_{cs}^{cb}} \quad (26)$$

$$R_{dist,actual}^{t,n} = R_{dist,ref}^{t,n}[1 + \alpha_{copper}(T_{actual} - T_{ref})] \quad (27)$$

$$P_{dl,es}^{t} = \sum_{n \in N} 3R_{dist,actual}^{t,n} I_{dist}^{t,n\ 2} \quad (28)$$

$$I_{dist}^{t,n} = \frac{P_{dist}^{t,n}}{V_{dist}^{t,n}} \quad (29)$$

$$P_{res,exc,pp}^{t} = P_{eg,es}^{t} - P_{dist,es}^{t} \quad (30)$$

$$P_{uel,es,ppp}^{t} = P_{uel,es,bpp}^{t} - (P_{dist,es}^{t} - P_{dl,es}^{t}) \quad (31)$$

$$DL\ (\%) = 100 \frac{P_{dl,es}^{t}}{P_{dist,es}^{t}} \quad (32)$$

11.6 Appendix VI: Model assumptions and parameters for Mongolian mobile micro-grids case study from chapter 6

Please refer to the nomenclature table presented at the beginning of this thesis for the definition and respective reference for each of these variables/ parameters.

Table 15: Parameters and assumptions used to build model in chapter 6.

Technical Parameters and Values/Ranges

T	30 years / 360 months	K_{eh}	1.52 USD/W
λ	5%	K_{ba}	0.7 USD/ Wh
η_{inv}	90%	α_{MS1}	0.95
η_{batt}	75%	α_{MS2}	0.85
$\eta_{coal\ stove}$	25%	$\varepsilon_{pe,avg}$	0.769
μ_{coal}	14.6 MJ/kg	k^{salv}	0.7
$A_{cs,cb}$	1.5 mm^2	${L_{pv}}^{max}$	25 years
ρ_{cb}	1.678x10^{-8} Ωm	${L_{wind}}^{max}$	20 years
T_{ref}	20 °C	${L_{pe}}^{max}$	15 years
$\alpha_{cb,adj}$	0.393 % / °C	${L_{eh}}^{max}$	13 years
EF_{grid}	0.711 tCO$_2$ eq./ MWh	c_{coal}	40 USD/ t
EF_{coal}	1.37 tCO$_2$ eq./ 1000 kg	$c_{om,pv}$	0.0048 $/kW/month
$\theta_{UB\ Grid}$	2 kW	$c_{om,wind}$	0.0032 $/kW/month
A_{grid}	1	$c_{om,pe}$	0.0001 $/kW/month
B_{grid}	9	c_{ul}	0.3417 USD/kWh
c_{cc}	0.04 USD/ kWh	c_{grid}	0.041 USD/kWh
c_{ic}	5.2 USD/m	p_{grid}	0.17 USD/kWh
$c_{pe,avg}$	0.352 USD/W		
K_{pv}	2.64 USD/W		
K_{wind}	1.91 USD/W		

GBM Model(s) Parameters

$\mu_{migration}$	0.5%
$\sigma_{migration}$	5%

μ_{ED}	3.14%
σ_{ED}	15%
μ_{HD}	1%
σ_{HD}	6.5%
σ_{PV}	19.3%
σ_{wind}	14.2%

11.7 Appendix VII: Hyperparameters used for ACER agent training in chapter 6

ACER Agent Hyperparameters Optimized via Optuna (not stated in Chapter 6)

n_{steps}	360	q_{coeff}	0.5
λ	0.99	Entropy factor	0.13
clip	10	Exploration final ϵ	0.1
Batch size	32	η	0.00009
ϵ_r	0.0005	α_r	0.99
Buffer size	15000	Replay ratio	7
Replay starts	4000	Correction term	10.0
α	0.95	Δ_{KL}	0.6

11.8 Appendix VIII: Hyperparameters used for PPO agent training in chapter 7

PPO Agent Hyperparameters Optimized via Optuna (not stated in Chapter 7)

n_{steps}	360	VF_{coeff}	0.5
λ	0.99	Entropy factor	0.05
Clip range	0.2	$Grad_{norm}^{max}$	0.3
Batch size	32	η	0.000025
Lam	0.85	Minibatches	4
n_{epochs}	4	Clip range	0.2

11.9 Appendix IX: Hyperparameters used for IQN agent training in chapter 8

IQN Agent Hyperparameters Optimized via Optuna

η	0.00002	Quantile Samples (N, N', K_τ)	8
λ	0.99	$CVaR_\alpha$	[0.25, 0.5, 1.0]
Exploration initial ϵ	1.0	Exploration final ϵ	0.03
Batch size	32	Network update	30000
k	10	Learning starts	90,000
ϵ_{PE}	0.35	Memory sampling	Every 4 steps