Exploratory Modelling and Reinforcement Learning of Merit-Order Electricity Markets

Marco Marasco

Supervisors:
Fjalar de Haan
Nir Lipovetzky
Enayat A. Moallemi
Angela M. Rojas-Arevalo

Master of Computer Science
School of Computing and Information Systems
The University of Melbourne

Abstract

This work explores the use of Reinforcement Learning as a tool for designing policies for systems under uncertainty. Our research investigates the efficacy of Reinforcement Learning to design policies, and how Dynamic Adaptive Policy Pathways (DAPP), can improve the quality of Reinforcement Learning derived policies in uncertain systems. The Victorian electricity market is used as a case study, where policies have been designed to support the transition to an environmentally sustainable future. A novel integration of the DAPP framework into Reinforcement Learning algorithms is proposed, to bolster the efficacy and robustness of Reinforcement Learning derived policies. Experimentation is also performed to better understand the Multi-Objective Evolutionary Algorithms (MOEAs) used by the DAPP framework to computationally design its policies. Our discussion on MOEAs evaluates what strengths they provide the DAPP framework for developing robust policies, and their implications for our proposed DAPP-Reinforcement Learning method.

A comparative analysis is conducted on the quality of policies designed using only Reinforcement Learning techniques, compared to policies designed using our DAPP-Reinforcement Learning method, in addition to various baseline policies. Our results show policies designed by the DAPP-Reinforcement Learning method on average increase Victoria's renewable electricity utilisation by 23%, and decrease household greenhouse gas emissions by 28%, when compared to the policies derived via only Reinforcement Learning algorithms. Through critical analysis of the results, this work conveys how the strengths of the DAPP framework can be combined with Reinforcement Learning to develop more robust policies for systems under uncertainty.

Declaration

I certify that:

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- the thesis is $28,544$ words in length (excluding text in images, table, bibliographies and appendices).

Signed,

Marco Marasco

October, 2021

Acknowledgements

It is not often a student has the opportunity to publicly express their gratitude to those who have supported them throughout their studies. In truth, I would need a second thesis to address everyone who made this project possible.

To Nir Lipovetzky, Angela Rojas-Arevalo, Fjalar de Haan, and Enayat Moallemi - what a remarkable journey this has been. I feel truly fortunate to have completed my thesis with not one, two, or even three, but four exceptional supervisors. Through what has been a trying year, your support and expertise have kept me motivated to pursue the highest quality of research. I endlessly appreciate your time, efforts, and involvement in my project.

I would like to give a special mention to Angela, without your work, *none* of this would have been possible. Thank you for allowing me to extend and explore Gr4sp, and I hope my project becomes one of many based on your work.

Lastly, to my family, partner, and friends - thank you for all the support and tortellini you have given me this year. I look forward to finally leaving my desk and spending time with you all.

Contents

1 Introduction 9
1.1 Background . . . 9
1.2 Research Questions . . . 11

2 Literature Review 12
2.1 Introduction . . . 12
2.2 Exploratory Modelling . . . 13
2.3 Reinforcement Learning . . . 17
2.4 Conclusion . . . 19

3 Research Methodology 20
3.1 Integration of Exploratory Modelling . . . 20
3.2 National Electricity Market . . . 22
3.3 Formalisation of the Adaptive Policy Problem . . . 23
3.4 Reinforcement Learning Agent . . . 24
3.5 Comparative Analysis . . . 25

4 Gr4sp Simulation Engine 26
4.1 Overview . . . 26
4.2 Merit-Order Electricity Markets . . . 27
4.3 Performance Indicators . . . 29
4.4 Policy Actions . . . 31
4.5 BAU Scenario . . . 33

5 Model Preparation 34
5.1 Python GSE . . . 34
5.2 Alternative Bidding Windows . . . 35
5.3 Historical Validation . . . 37
5.4 Selection of Bidding Window . . . 43

6 Augmenting Uncertainty 44
6.1 Sensitivity Analysis . . . 44
6.2 Identifying Important Uncertainties . . . 46

7 Constructing the Metro Map 49
7.1 Multi-Objective Robust Optimisation . . . 49
7.2 Experimental Setup . . . 52
7.3 MOEA Robust Optimisation Results . . . 60

7.4 Discussion 69
7.5 Constructing the Metro Map 70

8 Intelligent Policy-Makers 73
8.1 Markov Decision Processes 73
8.2 State & Action Spaces 74
8.3 Deep Q-Learning 75
8.4 Reward Function Design 76
8.5 Training For Deep Uncertainty 81
8.6 Evaluation 85
8.7 Discussion 91

9 Conclusion 93
9.1 Future Work 93
9.2 Concluding Remarks 95

Appendices 108

A GSE Appendix 109
A.1 GSE Uncertain Parameters 109
A.2 GSE Output Values 114

B Model Preparation Appendix 117
B.1 Forecast Data Smoothing Algorithms 117
B.2 Historical Validation Results 118

C Augmenting Uncertainty Appendix 122
C.1 Latin Hypercube Sampling 122
C.2 Sobol Indices 123
C.3 Sobol Indices Results 125
C.4 Ordered Uncertain Parameters 130
C.5 Factor Fixing Algorithm 132

D Metro Map Appendix 133
D.1 Stability Tests 133
D.2 ϵ-NSGAII Hyperparameters 136
D.3 Robustness Metrics Results 136

E Reinforcement Learning Appendix 138
E.1 Performance Indicator Mass Experiments 138
E.2 Reward Function Evaluation Results 138
E.3 Agent Training Data 140
E.4 Agent Evaluation results 141

List of Figures

2.1 Simple Metro Map 16

4.1 LCOE Examples 29

5.1 GHGE Levels Historical Validation 38
5.2 Renewable Market Share Historical Validation 39
5.3 Wholesale Prices Historical Validation 40
5.4 Tariff Prices Historical Validation 42

6.1 Pearson Correlation Per Performance Indicator 47

7.1 MOEA ϵ-Dominance Example 51
7.2 Robustness Metric Risk Table 54
7.3 Stability Test Example 57
7.4 Population Sizes 61
7.5 Termination 62
7.6 KDE Plots 64
7.7 Normalised Parallel Plot 65
7.8 Box Plots - Environmental Indicators 66
7.9 Box Plots - Economic Indicators 68
7.10 Four Simple Policy Pathways. 70
7.11 Simple Metro Map 70
7.12 Final Metro Map 71

8.1 Real-Time Performance Indicator Plots 79
8.2 KDE Plots 86
8.3 Normalised Parallel Plot 87
8.4 Box Plots - Environmental Indicators 88
8.5 Box Plots - Economic Indicators 89

D.1 Stability Test - GHGE Levels 133
D.2 Stability Test - Renewable Market Share 134
D.3 Stability Test - Wholesale Prices 134
D.4 Stability Test - Tariff Prices 135
D.5 Stability Test - Unmet Demand Days 135

E.1 Reward Function Evaluation Training Mean Rewards 139
E.2 Final Agent Training Graphs 140

List of Tables

4.1 Policy Actions List 32

5.1 Java vs. Python GSE Runtime Durations 34
5.2 Bidding Window Runtime Durations 37
5.3 GHGE Levels Validation Statistics 38
5.4 Renewable Market Share Validation Statistics 40
5.5 Wholesale Prices Validation Statistics 41
5.6 Tariff Prices Validation Statistics 43

7.1 ϵ-NSGAII Runtime Durations 60

8.1 Reward Function Policy Actions 80
8.2 Reward Function Average Performance Indicator Values 80
8.3 Reward Function Average Performance Indicator Baseline Change 80
8.4 RL Evaluation Acronyms 85

A.1 GSE Uncertain Parameters 109
A.2 GSE Output Values 114

B.1 Historical Validation - GHGE Levels 118
B.2 Historical Validation - Renewable Market Share 119
B.3 Historical Validation - Wholesale Prices 120
B.4 Historical Validation - Tariff Prices 121

C.1 Sobol Indices - GHGE Levels 125
C.2 Sobol Indices - Renewable Market Share 126
C.3 Sobol Indices - Wholesale Prices 127
C.4 Sobol Indices - Tariff Prices 128
C.5 Sobol Indices - Unmet Demand Days 129
C.6 Factor Fixing Pearson Correlations 130
C.7 Ordered Uncertainties 131
C.8 Final 12 Uncertain Parameters 132

D.1 ϵ-NSGAII Hyperparameters 136
D.2 Robustness Metric Results - GHGE Levels 136
D.3 Robustness Metric Results - Renewable Market Share 136
D.4 Robustness Metric Results - Wholesale Prices 137
D.5 Robustness Metric Results - Tariff Prices 137
D.6 Robustness Metric Results - Unmet Demand Days 137

E.1 Mass Experiments 138
E.2 Reward Function Training Duration 138
E.3 Final Agent Training Runtime 140
E.4 RL Agent Results - GHGE Levels 141
E.5 RL Agent Results - Renewable Market Share 141
E.6 RL Agent Results - Wholesale Prices 141
E.7 RL Agent Results - Tariff Prices 142
E.8 RL Agent Results - Unmet Demand Days 142

Chapter 1

Introduction

1.1 Background

Propelled by the depletion of fossil fuel reserves and environmental activism, a global revolution in the electricity sector is beginning to take place [1]. A key component to this revolution, is the rapid elimination of fossil fuel dependency, and uptake of renewable electricity sources.

Such transitions, also known as Sustainability Transitions, involve the transformation of existing infrastructural, economic and social systems, to a state that promotes the use of sustainable electricity sources [2]. To encourage these transitions, governing bodies seek to design policies using socio-economic tools such as financial investments, subsidies, and tariffs to embrace renewable electricity sources in society [3]. Typically, governing bodies target shifts in an infrastructural or economic fashion, such as reducing costs for renewable electricity [4], however, it has been argued in order to truly see significant transformations, shifts in social conditions such as consumer behaviour, or social acceptance of renewable electricity sources are required [4].

The diversification of electricity sources increases the heterogeneity of the sector, causing inherent complexities and uncertainties of the electricity sector to increase [5]. In addition, electricity sources are heavily influenced by many uncertain factors, ranging from social, political, economic, and technological [6]. This presents a significant challenge for governing bodies to adapt operation and management policies to compensate, a problem described as the "wickedness" of public policy [7]. When devising policies using quantitative methods, governments seek evidence-based insights, often by developing computational models, which can be extremely challenging for real-world systems. These models are then used to establish a set of strategies and policies to drive and to support targeted outcomes [1].

Forming a single computational model to use as a foundation for making predictions cannot appropriately manage the "wickedness" in planning of electricity transitions [8]. It is based on historical trends, or a limited number of speculated future scenarios, resulting in devised plans to fit these limited trends or scenarios [9]. Such restrictions on the breadth of plans can lead to undesired outcomes, with a notable example in poorly estimating electricity consumption and demand as a basis for tax incentive for renewable electricity, resulting in an economic climate that stunted renewable electricity growth [10].

Exploratory Modelling (EM) has presented itself as a computational aid for broader investigation into the field of Sustainability Transitions [8], as well as being a well-known method for assisting policy design under uncertainty [11]. EM provides a means to systematically explore the consequences of different parametric and non-parametric uncertainties of computational models [8]. Despite the potential advances posed by EM for literature in Sustainability Transitions, it still is a relatively unexplored area, with few papers contributing to the space [12]. As computational power has increased over time [13], Machine Learning has made significant contributions to system modelling and predictions. Machine Learning algorithms are able to vastly increase the utility of historical data, even in systems with uncertain dynamics [14]. Within the plethora of Machine Learning algorithms, Reinforcement Learning (RL) has been shown to be suitable for the control and optimisation of real-world systems by using historical data and system simulations. [15].

RL techniques have been used to model and mimic human policy design in complex environments [16], highlighting a potential use of RL in further assisting human policy-makers, particularly, how RL can assist in the policy design process. There is a plentiful collection of RL and electricity market literature, with past papers seeking to use RL to maximise profits for participants in electricity markets [13]. Despite this, there exists a gap in investigation of how RL can be used to regulate an electricity market [13], and further, how RL can be used to influence the market to promote renewable electricity sources.

Within the literature of Sustainability Transitions modelling, EM is a prevailing computational technique [17] for assisting in the policy design process, as it effectively produces insights to assist policy design by enumerating over possible futures of a given model [9]. Conversely, RL seeks to directly learn a policy [16], by repeatedly interacting with a model to maximise some reward. Despite these differences, both techniques share a common domain, to assist in the exploration and development of solutions to managing complex models. The use cases of both EM and RL are still being widely explored, and notably, no efforts have been identified that explore if EM and RL are able to be utilised as complementary techniques within a wider scope of solving the problem of policy design.

1.2 Research Questions

This project intends to explore the design of policies for Sustainability Transitions in electricity markets. Specifically, this project aims to address the gap in integration between EM and RL, by evaluating how RL agents can be used with EM techniques in the process of policy-making under deep uncertainty, using a case study of the Victorian energy market. A key concept for the co-operation between EM and RL is the policy pathway: the concatenation of multiple policy actions over time, where policy actions may be introducing new taxes, laws, or subsiding certain technologies. By employing EM techniques to produce a suite of optimised policy pathways as the action space for the agent, this study seeks to assess whether pre-processed optimised policy pathways can improve the quality of an RL agent, by providing a guide on what actions to take to lead to a desirable outcome. By integrating EM and RL, this study aims to provide a foundational piece of literature for the combination of EM and RL, and motivate further research into their integration.

Two motivational questions have been defined to guide the research aims of this study:

- **RQ1:** How can Reinforcement Learning algorithms regulate an electricity market modelled as a Markov Decision Process, to support the policy design for transitioning to sustainable electricity sources?
- **RQ2:** Can Exploratory Modelling combined with Multi-Objective Evolutionary Algorithms improve the quality of Reinforcement Learning derived adaptive policies in deeply uncertain systems?

This thesis is structured as follows. First, a review is completed on the current state of EM and RL in published literature (chapter 2), followed by a research plan (chapter 3), with discussion of the experimental process, and how the results will be analysed. Discussion on the computational model and its preparation used for answering the research questions is conducted in chapters 4-6. The experimentation and analysis of the novel integration is presented in chapters 7 and 8. A critical evaluation of this study's findings and concluding remarks are detailed in chapter 9.

Chapter 2

Literature Review

2.1 Introduction

Traditionally, policy-makers in complex real-world domains such as water management and transportation assume that the future can be predicted [18]. However, problems currently faced by policy-makers are increasingly defined by uncertainties about the future that cannot be reduced by gathering more information and are not statistical in nature [19].

These uncertainties may encompass a myriad of external factors, such as climate change, population growth, new technologies, economic developments, and their impacts, with such degrees of uncertainty having been defined as deep uncertainty. Formally, Maier *et al.* [20] described deep uncertainty as a condition when policy-makers are not able to agree upon, or do not know 1) appropriate computational models to describe a system's dynamics, 2) the probability distribution to capture the possible variation about certain parameters in the model, 3) or how to measure the utility of model outcomes.

A key challenge encountered by policy-makers faced with deep uncertainty, is if the future does turns out to be different from the predicted futures, policies are susceptible to failure [21]. As the future unfolds, any phenomena encountered that were not considered in the original policy design are compensated ad-hoc, a cumbersome and undesirable approach to dealing with unforeseen conditions [22]. Policies that are able to adapt to changing and unforeseen conditions are well suited to planning under deep uncertainty [23]. A policy planning paradigm that holds this value has emerged, known as adaptive planning. Adaptive plans are developed with the recognition that when faced with uncertainty, one needs to design policy plans to compensate for unforeseen real-world phenomena [24, 25]. Adaptive plans are designed to deal with contingencies as new data and information presents itself, mitigating the ad-hoc nature of making policy changes in traditional policy planning [22].

In the field of Sustainability Transitions, a research area frequently facing deep uncertainties, EM has emerged as an over-arching framework for assisting in the development of adaptive policy-making, and has received wide adoption into Sustainability Transitions research [22, 23]. Conversely, the rise of Machine Learning, and the emergence of RL has reached a wide range of different research domains, and have gathered the attention of researchers as a framework for developing policies in complex systems [26].

2.2 Exploratory Modelling

Developed at the RAND corporation [27] in 1992, Exploratory Modelling (EM) defines a process for how traditional computational models can be employed to support policy-making for systems with high complexity, and deep uncertainty. Central to EM is the ideal that any formal model developed is not an effective predictive model, and to discard any beliefs that a single model is sufficient. Instead, EM sees a formal model as a foundation to facilitate exploring over the range of possible values for uncertain aspects in the model, by computing vast numbers of computational experiments over the uncertain values.

Bankes proposed one of the uses of EM, known as data-driven EM, which pertains to seek out phenomena of interest within the system, such as what conditions lead to a specific outcome [27]. An example of data-driven is using EM to learn what circumstances lead to a worst-case outcome (worst-case scenario discovery [28]).

EM techniques have been widely employed in research as a support for policy-making in deeply uncertain real-world systems [29]. Its key strength to promote and guide the exploration of a deeply uncertain model has made it alluring to researchers in policy-design within complex real-world systems [29]. An example by Watson *et al.* [30], employed an EM technique, known as Scenario Discovery to conduct a vulnerability assessment as an analytical component of identifying preferred policies in many-objective optimisation problems.

The integration of EM has introduced novel methods to support policy-design, divided into two frameworks [31]. The first, adaptation frameworks, whose primary focus is to promote flexibility in policy design, and support the development of adaptive policies. The second, robustness frameworks, that principally support the design of static policies that operate desirably over a wide possibility of uncertain outcomes. The literature of several of these techniques have been reviewed and is presented over the following sections to further understand how EM has been used for adaptive policy-making under deep uncertainty.

2.2.1 Dynamic Adaptive Planning

Dynamic Adaptive Planning (DAP), designed by Walker *et al.* [24], is a foundational technique to the problem of adaptive policy design under deep uncertainty, and considered to be an adaptation framework. The ethos of this policy design approach explicitly considers policies will need to adapt to compensate for how the future unfolds.

DAP has been highly adopted in the literature as a supportive framework for policy-making in deeply uncertain environments, and operates at a high-level as follows. After developing a policy for a problem, researchers assess the quality of the policy via EM, conducting large-scale computational evaluations of potential future outcomes, and recording the policy's impact, particularly noting what conditions caused the policy to fail (data-driven EM). The policy is then updated to compensate, and the cycle is repeated until a policy of a desired quality is reached. Examples of DAP include planning for congestion road pricing [32], rail transport organisation [32], and urban transport infrastructures [33].

Despite its integration in academia, a reason for why DAP and the wider concept of adaptive plans had not been widely integrated in real-world practice was due to a lack of testing of the validity and applicability of these new planning approaches [34, 35]. Kwakkel *et al.* [36] sought to address this issue by assessing the efficacy of a DAP approach for guiding the long-term development of infrastructure of the Schiphol Airport in the Netherlands. This particular case study faced deep uncertainty including future travel demands, population growth, political and economic climates, and changing climate conditions.

The authors created a traditional static policy, and an adaptive policy derived via the DAP methodology. By utilising EM to assist the adaptive policy to explore over the possible futures of the airport, they were able to demonstrate the potential implications of actions made by the adaptive policy, and were able to change according to this feedback. The results indicated that the adaptive policy provided more favourable outcomes in a wider spectrum of the possible future scenarios, helping to cement the applicability of DAP for real-world policy-making.

DAP is one of the core concepts supporting the directions of this study. It has demonstrated its applicability in real world systems [32, 33], and provided a guiding framework for further research into adaptive planning. It does however, rely solely on human policy-makers to design, identify, and construct all components of the policy. Particularly, through each iteration there is a human choice on how to change the adaptive policies for a wider range of favourable outcomes. An alternate policy-making technique exists, that employs EM to computationally optimise each iteration of the policy-making process, known as Robust Decision Making.

2.2.2 Robust Decision-Making

Robust Decision Making (RDM), is a methodology that was designed to restructure the role of computational models and historical data for policy-making in deep uncertainty, resulting in static robust plans [37]. It combines a policy design techniques known as Assumption-Based Planning [38], with EM to pressure test policies over a plethora of future scenarios. A robust policy is one that performs well, compared to the alternatives, over a wide range of plausible futures [9, 39]. RDM's application has extended to develop management policies for many deeply uncertain real-world systems, including urban water infrastructure [40–42], flood risk [43], and electricity resources [44].

Using the ideology of EM, RDM explores different possible future scenarios to provide a platform for rigorous investigation of computational models. By doing so, RDM assists in identifying a plan that performs adequately over a wide range of potential futures, supporting the notion of robustness in an identified plan. RDM is often used in combination with EM to help enumerate the plethora of potential future scenarios, to assist in identifying and mitigating the conditions that would cause a given policy to fail, in line with Bankes' data-driven use for EM [9, 41].

RDM's strength lies in its optimisation process to assist in developing a single robust plan that will perform adequately in a wider range of scenarios. This optimisation does not come without cost, robustness itself is not a definitive metric, there are multiple ways of measuring if a plan is robust, with each measure having its own trade-offs, and a high computational cost [45]. Regarding the static nature of RDM, in the realm of deep uncertainty, non-static plans are susceptible to failure [46]. This is a fundamental manner to which RDM differs from DAP, and DAPP (section 2.2.4). It does not consider any adaption of the plan over time, whereas that is the crux of DAP and DAPP. This is a comparative weakness for RDM's applicability to deeply uncertain problems.

2.2.3 Many Objective Robust Decision-Making

As previously stated, computing policies via RDM is computationally too expensive, with optimal solutions often becoming intractable for many-objective problems in complex systems [47]. Multiple different solutions have been developed to translate multi-objective problems to single-objective, such as the utility function method, weighted sum approach, goal programming, and lexicographic method [48–51]. These approaches require input for weighting the different objectives against each other, causing concerns for bias to be present [50].

Kasprzyk *et al.* [47] utilised Multi-Objective Evolutionary Algorithms (MOEA) to assist in the computational process for RDM to allow for a new method of identifying robust policies for systems under deep uncertainty. Named Many-Objective Robust Decision Making (MORDM), it allowed policy-makers to generate a suite of policies to assist policy-makers to select a policy that performed well under a wide range of future scenarios. Using a case study for water management in the Lower Rio Grande Valley in Texas, USA, their work demonstrated MOEAs were able to produce computationally tractable sets of policies, each optimised to operate robustly.

Kasprzyk *et al.*'s direction into using MOEAs presented a new way for how RDM could be incorporated into policy-making, as its potential intractability was mitigated. Hamarat *et al.* [52] utilised MORDM, in conjunction with past DAP literature, to marry the two policy-making frameworks, aptly titled Adaptive Robust Design (ARD). Using the European Union's electricity market carbon emission scheme (ETS) [53], Hamarat *et al.* demonstrated how ARD was able to produce more desirable outcomes over the potential futures of the EU market, compared to the existing ETS policy, and a policy derived purely by the DAP framework. Their key finding was that the robust optimisation generated by MORDM allowed for a wider range of favourable outcomes, compared to non-robust policies.

Hamarat *et al.*'s case study demonstrates the aims of this research to explore adaptive policies in transitions of electricity sectors are well founded in literature, and also provides a strong basis for the gaps in research to be addressed. ARD generates a collection of robustly optimised adaptive policies, which are used by policy-makers to interpret the potential future policy pathways they may follow to reach some outcome. Each policy that ARD generates is distinct, there is no option within this framework to choose one policy, and then change it later. This lock to a given plan can be mitigated through an emerging policy planning paradigm, Dynamic Adaptive Policy Pathways.

2.2.4 Dynamic Adaptive Policy Pathways

Dynamic Adaptive Policy Pathways (DAPP) [54] combines two bodies of literature on planning under deep uncertainty: adaptive policy-making and adaptation pathways [24, 55, 56]. DAPP supports the exploration of a wide variety of relevant uncertainties in a dynamic way, connecting short-term targets and long-term goals, by identifying short-term actions while keeping options open for the future.

DAPP is described as a more complex and rigorous framework to assist the formation of high quality adaptive strategies for human planners facing deep uncertainty [54]. DAPP was designed with the realities of government-level management in mind, particularly, the notion of unstable political environments, and the possibility of different people managing a policy. It supports this instability by allowing policy-makers to understand how long different short-term decisions can be postponed before jeopardising long-term goals, allowing for more flexibility in short-term actions.

Arguably, the greatest contribution from DAPP was the introduction of the *Metro Map* (figure 2.1), a visual representation of the different potential policy pathways, and how they are able to interact. This map is intended to be a guide for policy-makers to navigate, and to understand how they may adapt their plans as time progresses.

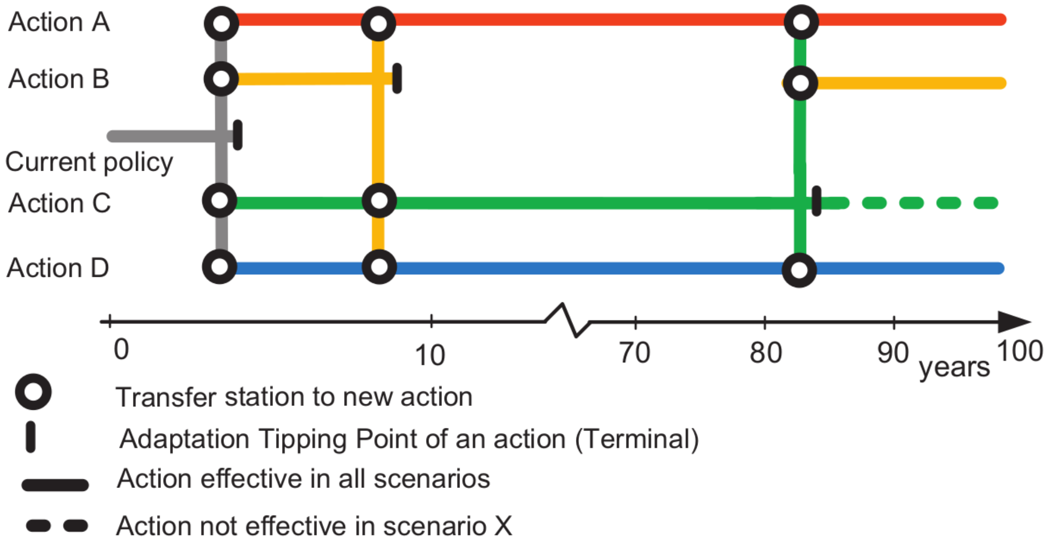

Action A
Action B
Current policy
Action C
Action D
0
10
70
80
90
years
100
Transfer station to new action
Adaptation Tipping Point of an action (Terminal)
Action effective in all scenarios
Action not effective in scenario X

Figure 2.1: A simple *Metro Map* generated by DAPP [54].

Whilst the map is highly useful as a final product, its creation requires addressing of the multiplicity of potential policy actions combinations, or visually, traversing the final *Metro Map*. The *Metro Map* represents policy pathways to be followed over time, with the vertical axis representing the current policy action taken. Transfer stations exist between policy actions to represent the opportunity of taking a new policy action, and simultaneously retiring the current policy action.

There are two main challenges for identifying the policy pathways to use for constructing a *Metro Map*: selecting pathways from a potentially infinite collection of possible pathways; and ensuring identified pathways are robust to the uncertainties in the system. A solution to the first challenge was addressed by Kwakkel *et al.* [57], based off Kasprzyk *et al.* [47] work using MOEAs, Kwakkel *et al.* integrated MOEAs to develop a set of non-dominated policy pathways, then aggregated them into a single map. Notably, this is an effective aggregation of pathways developed by Hamarat *et al.*'s MORDM [52], addressing the weakness in MORDM on how to combine and change between different policy pathways found by MORDM.

The results of Kwakkel *et al.* [57] work are promising for the direction of this research. DAPP was shown to not only incorporate the fundamental adaptive policy techniques from DAP, but also included RDM and the extension of MORDM to incorporate the notion of robustness into planning. However, as noted in the concluding remarks of [57], like all methodologies compared in this review, it does not provide guidance on how to best utilise its deliverable map. Accordingly, a challenge of DAPP, deciding which pathway to traverse, and methodologies to assist in this final decision process, are still an open problem for DAPP, and the wider community of policy-making under deep uncertainty.

2.3 Reinforcement Learning

Reinforcement Learning (RL), a Machine Learning technique, is concerned with designing intelligent agents to learn what actions to take in a given environment to maximise a measurable outcome (reward). RL is able to elicit system dynamics knowledge from historical data by interacting on a continual basis with an environment, effectively, learning by doing. It is able to mitigate dependencies on traditional computational models by learning a proxy model [16], or via a process known as Batch RL [58].

This approach to learning by doing has highlighted RL's ability to imitate human-level decision capabilities, spurring research into its applicability in different sectors. Prominent sectors that have adopted RL include autonomous robotics, communications networking, and biological data manipulation [59–61].

2.3.1 Electricity Markets

Due to ever increasing complexities ranging from environmental to political factors, and the resultant uncertainties from a diverse system, a systematic shift in the archetype of electricity management problems has been suggested necessary to keep up with system complexities [62], prompting a surge in research for new methodologies, notably, the adoption of RL into the domain space.

Specifically regarding the open electricity market, these rising inherent complexities have motivated exploration into more sophisticated approaches to market participation (bidding) strategies, to maximise profit for electricity generators in these deeply uncertain environments. RL has been strongly embraced into research for the management and optimisation of market participants [13, 63–65].

A key example for the basis of this study's research aims, Subramanian *et al.* [66], focused on the problem of maximising the profit of a given generator participating in a single-sided electricity market. By utilising historical logs to model the Australian National Energy Market, they were able to leverage the model to train an RL agent to learn bidding strategies to maximise profit via the RL technique known as Deep Q-Learning (DQN) [67]. Subramanian *et al.* utilised DQN for its ability to handle models with continuous state spaces. For a given generator bidding in the electricity market, they were able to demonstrate their RL agent received higher returns over a given time period, compared to the generator's historical values. Subramanian *et al.*'s approach to modelling the Australian National Energy Market provides a sound basis of literature for the model required for this study.

Despite these advances into RL applications for electricity markets, there is an abundant lack into investigation of its application to a market regulator. Past papers are concerned with maximising profits for generators [13], and do not consider how RL can assist in market regulation in transitions to renewable electricity sources, a gap intended to be addressed in this study.

2.3.2 Adaptive Policies

In stochastic systems, such as electricity markets, adaptive policy problems can be mathematically modelled using Markov Decision Processes (MDP) [68, 69]. MDPs are stochastic processes that satisfy the Markov property (the future depends only on the present [70]), and assign different rewards/costs for a system transitioning from one state to another [71]. MDPs are of significant value for RL approaches to adaptive policy-making, as MDPs are an extremely common way to frame RL problems [16].

Designing an MDP to model a system with uncertainty diverges from MDP modelling approaches where the system is fully observable, due to values of model parameters, or even the true dynamics of the model being uncertain to the policy-maker [72]. Within MDP literature, MDPs developed within partially observable systems are known as Partially-Observable MDPs (POMDP) [73], where probability distributions of uncertainties in the system are modelled directly into the state of the POMDP. Chades *et al.* [74] argued when designing models under deep uncertainty, the uncertainty is around the value of parameters of the model, known as parameter uncertainty in the literature for MDP adaptive management [73].

In MDP modelled adaptive management problems, the task to resolve parameter uncertainty is to manage the system while simultaneously learning the value of the parameter to improve future management decisions. Walters *et al.* [75] developed one of the first, and widely adopted techniques for parameter-uncertain models, by taking advantage of updating prior and posterior beliefs of uncertain parameters in managing fisheries. This approach utilised a normal distribution to determine uncertain parameter values, in combination with environment variations, presenting a posterior normal distribution for uncertain parameters. The advantage of this approach was the ability to design a closed-form posterior distribution, removing the computational need for simulation methods. Their method has been adopted in multiple studies pertaining to adaptive policies in deep uncertainty [76–78].

Abstracting from the modelling process, further literature has been reviewed pertaining to the specific application of RL in developing adaptive policies. Huang *et al.* [79] advocate DQN as a RL algorithm for efficiently solving an adaptive policy problem for a smart-grid, a new electricity network framework designed to take advantage of computer assisted monitoring [80]. The result of Huang *et al.*'s study was a novel control system, that leveraged the feature extraction capabilities of DQN, and its functionality to generalise complex, and uncertain systems that were modelled as MDPs. Huang *et al.*'s research into using RL for electricity management is not isolated, with multiple other papers [81–83] employing DQN for adaptive management, further demonstrated the utility of DQN for adaptive management problems in deep uncertainty.

Zhang *et al.* [84] alternatively approached their adaptive management using the Deep Deterministic Policy Gradient (DDPG) algorithm to train their RL agent. In Zhang *et al.*'s study, they sought to develop a data-driven approach for the adaptive management for a wind turbine, a problem plagued with many environmental and economical uncertainties. Compared to DQN as employed by [79], DDPG is more suited to problems with large action spaces [84].

The reviewed literature pertaining to adaptive management for RL has highlighted the two gaps to be addressed in this study. Firstly, the lack of research into RL based adaptive management for electricity market regulation, and a further extension, how a RL based adaptive policy can regulate the market to support a transition to renewable electricity sources. Further, the abundance of papers demonstrating the efficacy of modelling these problems as MDPs and RL techniques for solving adaptive management problems reinforces this study's utilisation of RL to regulate the electricity market.

2.4 Conclusion

Regarding EM techniques, five extant approaches were identified, Dynamic Adaptive Planning, Robust Decision Making, Many-Objective Robust Decision Making, Adaptive Robust Design, and Dynamic Adaptive Policy Pathways. Dynamic Adaptive Planning, as well as Robust Decision Making both demonstrated their own individual merits, but these early techniques have been improved and extended to develop newer approaches such as Adaptive Robust Design and Dynamic Adaptive Policy Pathways. The incorporation of MOEAs to policy-making by [47] has provided a strong foundation into new approaches to developing policies, with the notable uptake in the DAPP process. MOEAs in combination with DAPP will assist in developing effective policies for the case study of this research.

For RL, the established literature around the electricity market for bidding strategies, as opposed to regulatory policies demonstrates a gap in research for investigating the potential use of RL for the case study in this paper. By reviewing past methods to design adaptive policies for deeply uncertain systems, this review learned of the employment of Markov Decision Process models, and their frequent use with RL for developing adaptive policies. This investigation into MDPs for adaptive policies will have a high utility when designing the models required for this study.

Chapter 3

Research Methodology

To effectively address the research questions, a detailed plan has been proposed that will guide this study, with the support and justification of previous literature. The experimentation will require a simulation model of the National Electricity Market (NEM), the specification of how to measure the NEM performance, and policy actions to influence the NEM. Once the simulation model has been prepared, a RL agent, and an EM assisted RL agent will be generated, and their efficacy will be tested and evaluated. This research methodology will outline how the electricity market will be modelled, the manner in which the adaptive-policy problem will be designed, the way EM will interact with RL, and how data will be generated and analysed to answer the research questions.

3.1 Integration of Exploratory Modelling

To address the research aims, specifically, how EM can aid RL, this study intends to explore the potential use of EM generated policy pathways. ARD and DAPP, both produce a suite of potential policy pathways for policy-makers to review, and use as a guide. The policy pathways ARD and DAPP produce are created via MOEAs, which explore the possible policy pathways (policy space), and optimise the returned pathways for robustness accordingly to some multi-objective goal. For policy-makers, this can be rationalised as pre-processing the space of all possible policy pathways, and removing undesirable pathways from the final consideration.

However, we postulate that this pre-processing of the policy pathway space, is analogous to pre-processing the action space for an RL agent. For an RL agent, the action space is equivalent to the policy pathway space for the EM frameworks: a concatenation of RL agent actions over time. By producing a suite of already optimised policy pathways as the action space for the agent, this study seeks to assess whether the pre-processed optimised policy pathways can improve the quality of an RL agent, by providing a guide on what actions to take to lead to a desirable outcome.

As identified in the chapter 2, each policy that ARD generates is distinct [52], there is no option within this framework to choose one policy pathway, then change later. The pre-processing provided by ARD simply reduces the action space to a set of distinct paths, and thus the agent's action space is to choose a path at the first timestep. After choosing a path, the agent is no longer required, and the policy actions in the pathway are executed according to the pathway.

Contrarily, DAPP addresses this single pathway lock-in by producing a *Metro Map* (figure 2.1) to interweave adaptive policy pathways where possible [57], providing greater flexibility and control for policy-makers to change pathways where permitted. As such, the DAPP pre-processes the action space to a more dynamic action space, where decisions and changes can be made along policy pathways. Translating this to the context of a RL agent, the *Metro Map* represents the possible sequence of actions RL agent can make, with the multiplicity of sequences made possible by different pathways in the *Metro Map*.

Roughly, the problem is reduced to navigating through the *Metro Map* that DAPP generates, in order to reach a final state that maximises performance indicators. From this comparison, this study investigates the use of DAPP to provide the most effective framework for assisting RL in generating an adaptive policy, as it can provide a non-deterministic policy pathway after the first timestep, allowing for more in-depth exploration into the use of EM in RL.

3.1.1 Implementation of Dynamic Adaptive Policy Pathways

This study will seek to follow the computational procedure for developing DAPP by Kwakkel *et al.* [57], proposed by the same authors of the original DAPP paper [54]. Kwakkel *et al.* outline the steps required to explore and optimise a set of policy pathways for an adaptive policy problem. In addition, they provide justification of the MOEA they employed, the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [85], which was cited to be well-known and applied to a wide range of problems [86].

Since the original paper by Kwakkel *et al.* [57], a new technological platform has been developed by Kwakkel *et al.* to further assist in the computation of EM. An application written in the Python programming language [87], titled the Exploratory Modelling and Analysis (EMA) Workbench, aims at providing support for performing computational EM with models developed in various modelling packages and environments. Since its inception [88], the workbench has been an integral part in many papers using EM [52, 54, 89, 90], and has proven to be an effective tool to assist in computational implementations. For the purposes of this study, the workbench also possesses features for experimenting with MOEA algorithms, providing a simple computational interface for developing the DAPP pathways.

3.2 National Electricity Market

The National Electricity Market (NEM) [91] is a wholesale electricity exchange for the five Australian states of Victoria, New South Wales, Queensland, South Australia, Tasmania, and the Australian Capital Territory. The market operates as a spot market, where electricity supply and demand is matched instantaneously through a centralised dispatch process.

Electricity generators submit bids to supply the NEM with specific quantities of electricity at a pre-determined price of their choosing. Every day, a generator submits a series of bids to the NEM, with each bid corresponding to the demand required in a given five-minute interval over the following twenty-four hours. Using all the received bids, the Australian Energy Market Operator (AEMO) determines which bids will be accepted to produce electricity, with the smallest price bids accepted first. This ordering of accepted bids is known as a merit-order. The NEM is designed to attain its demand requirements over the following twenty-four hours in the most cost-effective manner possible. Once enough bids have been cleared to meet demand requirements, AEMO provides each generator with a schedule for when they are permitted to generate electricity, according to the respective successful bids of each generator. For every five-minute interval, AEMO determines a uniform price, known as the dispatch price, according to the successful bids for that interval. Six dispatch prices are averaged every half hour to determine the spot price. The spot price is used as the basis for financial settlements for all generators in the market.

3.2.1 Modelling the National Electricity Market

A recent study by Rojas-Arevalo [92] has shown generators in the context of the NEM, follow bidding strategies defined in terms of their levelised cost of electricity, nameplate capacity, and other generator specific attributes, to a degree close enough to reproduce actual historic bidding trends. This study will incorporate a computational model of the Victorian electricity network within the NEM, developed by Rojas-Arevalo [92] for their investigation of socio-technical layouts of electricity systems, and their impacts on sustainability [93]. This model was generated using relevant NEM data obtained from a wide variety of government sources, including AEMO, the Australian Bureau of Statistics, and the Australian Bureau of Meteorology. The breadth of data sources provided Rojas-Arevalo a rich repository to create and model the NEM, whilst taking into account environmental and socio-economic factors.

Rojas-Arevalo utilised Agent Based Modelling (ABM) [94], a paradigm of computationally modelling for simulating interactions between different entities, with the aim to evaluate the effects of their interactions on the simulation environment as a whole. ABM provided Rojas-Arevalo with the ability to simulate participants in the NEM, and specify their attributes and behaviour using historical values for their generator specific attributes. In addition, the wider simulation program was designed with the ability to input all potential market parameters features, including those not known by generators, such as thresholds for policy actions, and impending tax adjustments.

This greater breadth of parameters provided the model with the capability to generate a wide variety of potential dynamics and outcomes. Accordingly, the complexity of such a model's behaviour can be effectively utilised to represent a deeply uncertain system, by varying the combination of input parameters. Rojas-Arevalo's formalisation of generator allows for effective forecasting of NEM electricity demands and behaviour, from 2020 to 2050. This will be of high utility for this study, as it will provide an environment to explore and design policies for future market conditions, through the use of both RL and EM techniques.

3.3 Formalisation of the Adaptive Policy Problem

Adaptive policy problems, as shown in the literature, can be mathematically modelled using Markov Decision Processes (MDP) [68, 69], a common way to frame problems to be solved by RL [16]. As learned in the review, many past studies have utilised MDPs to mathematically describe adaptive policy problems, and as such provide sufficient evidence MDPs are an appropriate representation for the adaptive policy problem. The output data from Rojas-Arevalo's simulation model will provide a reference set of data to model the state of the simulation system at a given time point. The formal description of the system state will be designed to provide sufficient statistics [73] for an RL algorithm to determine the best way to interact with the simulated system.

3.3.1 Policy Actions

As the case study seeks to simulate a real-world market, careful consideration needs to be made when determining what actions would be evaluated in the model. Key factors to consider for policy actions include feasibility, cost, applicability, and impact [95]. Previous real-world case studies have employed expert knowledge [54, 96] to effectively determine the best ways to interact with the simulation. An alternative method amongst the reviewed papers, was to use publicly available documentation from official government bodies, as well as independent government-commissioned recommendations for different approaches to interact with their system [26, 64, 66, 97]. Due to time constraints of this research body, this study will consult publicly available government and independent documentation pertaining to potential market mechanisms that can be applied to the generators to influence their behaviour. Several documents from AEMO, the Victorian state government, and the Australian Energy Regulator have been identified, providing a preliminary basis for policy action development. Examples from these documents include subsidising operational costs of generators, restricting a generator's ability to participate in the market, and carbon emission related taxes.

In respect of the ethos of Sustainability Transitions literature, the performance indicators of the MDP will be inspired by Rojas-Arevalo's study [92], specifically, greenhouse gas emission levels, the percentage of renewable electricity in the NEM, wholesale electricity prices, electricity tariff prices, and the number of unmet demand days (detailed in chapter 4). These will provide the means to measure an overall goal to minimise both environmental and economic impacts for any policies implemented, and support exploring a transition to a larger share of renewable electricity sources.

3.4 Reinforcement Learning Agent

3.4.1 Technological Platform

To maintain the same technological platform as the EMA Workbench, this study determined to use the Python library, OpenAI Gym [98]. OpenAI Gym possesses ready-made interfaces to easily integrate simulation environments with RL algorithms, exposing the necessary data for RL algorithms. Additionally, OpenAI Gym contains a wealth of community support and libraries that are designed to integrate, and promote RL research [99].

3.4.2 Proposed RL Algorithm

An integral part to the success of an RL agent is the algorithm used for training [16]. For the purposes of this study, any RL algorithm chosen must be able to handle environments with deep uncertainty. From the current knowledge obtained from reviewed literature in modelling electricity markets, from both EM and RL literature, this study reasonably estimates the state space of the final MDP will be continuous, but not necessarily the action space. Past RL papers in electricity market participation were consulted to guide this study towards the most appropriate algorithms, as they possessed relevant contextual issues of uncertainty and state/action space [13, 63–65, 79, 84].

Two different RL algorithms presented themselves as common approaches to RL in deep uncertainty, Deep Q-Learning (DQN), and Deep Deterministic Policy Gradient (DDPG). As stated in the review, both are able to operate over large state spaces, but differed where DQN was more suited to problems with a discrete action spaces [84]. As the timing of this study will not allow effective consultation with industry experts, a finite number of policy actions will be used to simplify the action design process. As a result, the DQN algorithm will be used for experimentation.

To implement this RL algorithm, an existing Python RL library, RLlib [100] will be used. RLlib possesses implementations of a wide variety of RL algorithms, including DQN, and native support to the OpenAI Gym API, further consolidating its utility for this study. By using a pre-implemented algorithm, the risk of incorrectly implementing an RL algorithm is reduced [100], and its public availability promotes reproduction and extension of the experimentation to be conducted.

3.5 Comparative Analysis

To understand the impact of the integration of EM into RL, and the resultant adaptive policies, a pure RL adaptive policy is required to compare and contrast against the policies derived by the DAPP/RL technique. For the electricity market simulation via the NEM model, two RL agents will be trained with the DQN algorithm, using the OpenAI Gym. The first agent, a pure RL agent, will be trained to develop a policy using the DQN algorithm, and will demonstrate the ability for RL to develop adaptive policies under deep uncertainty. The second agent, a DAPP/RL agent, will utilise the proposed EM technique, by pre-processing the action space of the agent via the DAPP methodology outlined in section 3.1. The pure RL agent, and the DAPP/RL agent respectively assist in answering research questions **RQ1** and **RQ2**.

In line with past studies assessing the quality of adaptive policies [52, 101], this study will utilise Exploratory Modelling and Analysis (EMA) as the methodology to test and analyse the quality of the adaptive policies. EMA assists policy analysis in deep uncertainty by exploring over the range of uncertainties in a given model, ranging from parametric, structural, and method uncertainties, using computational models for each scenario [102]. For this study, the model will be the NEM model (section 3.2.1), and the uncertainty is over the parameter uncertainty of Rojas-Arevalo's model, with different parameters passed as inputs to the model.

Using the EMA framework, both adaptive-policy agents will be run over numerous potential future scenarios, with variations generated by different parameters passed to the NEM simulation model to explore the model uncertainty. The runtime values of the performance indicators for each simulation and agent will be recorded, and used for analysis of the two adaptive-policies. This exploration of a wide variety of uncertainties will produce a vast dataset that will require further analysis in order to derive relevant information for comparing the quality of the two policies.

A comparative analysis on the range of outcomes will be completed, following the methodology by Kwakkel *et al.* [102]. Kwakkel *et al.* proposed to compare the quality of two policies by analysing the range and distribution of its performance indicators, arguing a smaller range and tighter distribution was more favourable due to its greater potential predictability for uncertain futures. EMA is supported by the EMA Workbench [88], and thus will be used in this study for generating and analysing experimental data. Once the data is generated, the results will be analysed, and with justification from the analysis, a final discussion will be made to formally address the results of the research aims of this study.

Chapter 4

Gr4sp Simulation Engine

To conduct their investigation of the Victorian electricity network within the National Electricity market, Rojas-Arevalo [92] developed the Gr4sp software package, a suite of tools for the experimentation and analysis of socio-technical layouts of electricity systems, and their impacts on sustainability [93]. A key component of the Gr4sp package is the Gr4sp Simulation Engine (*GSE*), a computational model used for simulating the Victorian electricity network, and is written in the Java programming language [103]. This chapter discusses the structure and dynamics of the *GSE*, as well as the performance indicators and policy actions that have been designed that will be used to assist answering the research questions of this study.

4.1 Overview

The *GSE* is made up of two main modules, a recursive Service Provision Module (SPM) structure, and a market simulator. The recursive SPM structure is an abstraction of how elements of the electricity system are organised, such that the recursive nature reflects the supply chain of the system. For example, electricity generators supply electricity to electricity retails, who supply electricity to consumers. The market simulator is an Agent-Based model (ABM) which has been designed to approximate the bidding mechanism within the NEM. The main components of the market simulator are electricity generators, who are each individual agents, with different attributes relating to their electricity capacity, fuel sources, and prices.

Prior to runtime, a set of input parameters are passed to the *GSE* to change the behaviour of the model. These include, forecast inflation rates, consumption levels, and prices of fuel sources for electricity generator (e.g. coal, gas). When using the *GSE* to explore simulation over future dates, these input parameters become assumptions of how the future may unfold. This key feature of the *GSE* to explore different assumptions of future behaviour has allowed [92] to conduct Exploratory Modelling techniques to gain insight into the possible trajectory of the Victorian division of the NEM under future uncertainty.

The input parameters can be considered as input uncertainties, where each parameter represents a social, economic, or technological aspect of the modelled NEM, that cannot be known in advance. In Exploratory Modelling literature, the set of input uncertainties passed at runtime to a model is known as a scenario. Citing Maier *et al.*'s definition of deep uncertainty [20] (see section 2.1), making predictions on future behaviour using Rojas-Arevalo's *GSE* can be considered a deeply uncertain problem, as the dynamics of the model are unknown when uncertain parameters are input at runtime. The full list of the 33 input parameters can be found in Appendix A.1. Throughout its execution, the *GSE* records a comprehensive yearly summary on the simulated behaviour, and attributes of the *GSE*, providing output values relating to electricity consumption, prices, and emissions throughout each execution of the *GSE*. The full list of output values can be found in Appendix A.2.

When designing policies for public systems, it is imperative to evaluate the social, economic, and temporal feasibility of any policy actions that are considered [95]. Specifically for electricity market regulation, policy actions should not be frequently implemented, and should operate on a larger, yearly scale. It was decided to implement new policy actions at the beginning of each Victorian state-level political cycle, of which the next cycles begins in 2022. Following the political cycle assists in ensuring actions are appropriately spaced, and strengthens this study's utility as a guide for future Victorian governments.

The operation of the *GSE* for this study will be completed as follows. Given a set of input uncertainties (scenario), the *GSE* will simulate the market from 01/01/2019, to 31/12/2050 (the final date the model is able to simulate). The model will initially run with no policy actions implemented for three years. This allows the input uncertainties to influence the market to a large enough degree that the state of the market at the end of 2021 will be distinct amongst all possible sets of input uncertainties. Every four years, beginning in 2022, a policy action will be implemented to affect the system, that will operate until the beginning of the next political cycle, and the final new policy action activated will take place in 2046. Although there is one more political cycle beginning in 2050, any new policy actions activated in that year will only have one year to influence the market's behaviour. Therefore, policy actions activated in 2046 will run for 5 years, until the final simulation date.

4.2 Merit-Order Electricity Markets

A key concept for understanding the *GSE* is how the market simulator operates, and the notion of bidding rounds. During the simulation time period, every 30-minutes the market simulator forecasts the amount of electricity required for the Victorian electricity system, known as the demand. Each electricity generator then submits a bid to the market to supply a specific amount of electricity at a price per electricity unit of their choosing. The market collects all bids, and orders them by price, lowest to highest (merit-order). The market operator first accepts the lowest priced bid, and subtracts the amount of electricity to be supplied by that bid from the total required demand for that 30-minute window. The next lowest priced bid is accepted, and that process is continued until the electricity supply meets the demand. Once demand has been met, the price of the last accepted bid (the LCOE of the last generator) is set as the wholesale price of electricity for that bidding round

A common financial metric for electricity generators is known as the levelised cost of electricity (LCOE) [104]. The LCOE determines the average price a generator needs to sell its electricity for, in order to meet the costs of building and operating over the expected lifetime of the generator. Rojas-Arevalo [92] was able to demonstrate how an electricity generator's LCOE can influence the generator's bidding strategy in the NEM, to a degree close enough to reproduce actual historical values. A generator's LCOE influences the price of electricity a generator submits in a bid to the market, and thus, influences the bidding power of the generator within the merit-order mechanism of the market. A high-level explanation of the LCOE for electricity generators is provided below.

Consider an electricity generator G, who is able to produce $C = 100MW$ of electricity, at a base price of $P = 20.00\$/MWh$, where the base price is the minimum price the generator is willing to sell electricity for. The generator is only able to operate at a maximum of $M = 80\%$ of its production capacity. The generator also records the number of times it has submitted a bid N, its historic revenue R, and a metric known as its historic capacity factor H. The historic capacity factor is used to adjust the LCOE of the generator to ensure the LCOE is set to cover the expected lifetime operating costs of generator G.

$$H = \frac{R}{\frac{C}{2} \cdot \frac{P}{M} \cdot N}$$

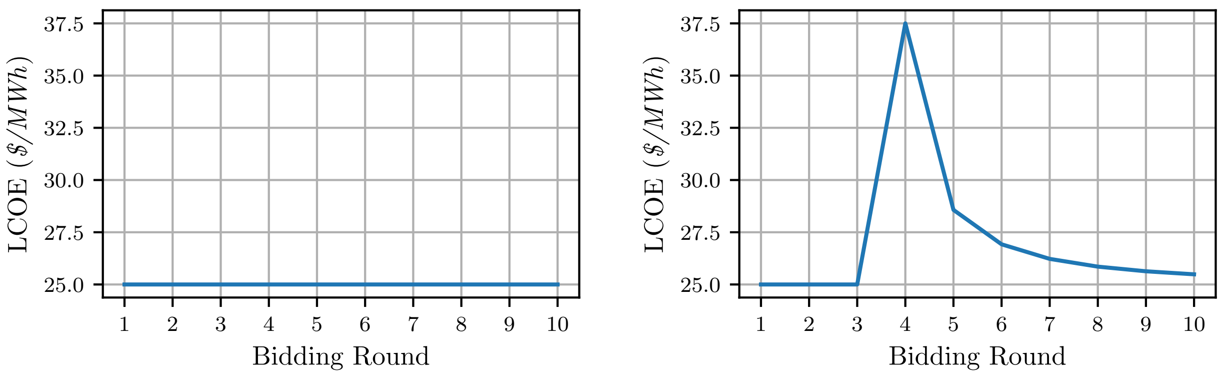
When a generator submits a bid to supply electricity in the GSE, the model assumes the generator supplies the maximum electricity it is able to, defined by $C \cdot M$, and the bid price is set by its LCOE, which is defined below. When the first bidding round begins in the model, as $H = 0$, the bid price is set to ${}^{P}/_{M}$.

$$LCOE = \begin{cases} \frac{P}{M}, & H = 0 \\ \frac{P}{H}, & H > 0 \end{cases}$$

Suppose generator G sets its LCOE to ${}^{P}/_{M} = 25.00\$/MWh$ in the first 30 minute $(0.5h)$ bidding round. In this round, it is able to generate and sell $C \cdot M \cdot 0.5h = 40MWh$ of electricity. Generator G's historic capacity factor is updated accordingly:

$$H = \frac{40MW \cdot 25.00\$/MWh}{\frac{100MW}{2} \cdot \frac{20.00\$/MWh}{0.8} \cdot 1} = 0.80$$

For the next bidding round, the price will be set to ${}^{P}/_{H} = {}^{20}/_{0.8} = 25.00\$$. Figure 4.1a demonstrates that so long as a generator's bids are successful, its LCOE will never change. In the event a generator's bid is rejected, their historical revenue, R, does not increase, which causes the historic capacity factor, H, to decrease as N is incremented without R increasing. Consider the LCOE curve in figure 4.1b, where generator G had an unsuccessful bid in the third bidding round. This causes the LCOE to increase from $25.00\$/MWh$ to $37.50\$/MWh$, as the generator needs to charge more to recover the losses from the unsuccessful bid.



(a) All successful bids.

(b) Failed bid in round 3.

Figure 4.1: LCOE illustrative examples.

As the merit-ordering of bids in the NEM is price based, an unsuccessful bid results in generators losing their bidding power, as their LCOE is increased. This relationship is essential for the design of policy actions for this study, as actions can be designed to impact the LCOE, and therefore, the bidding power of generators in order to manipulate the market.

4.3 Performance Indicators

Five performance indicators have been chosen, in accordance with the ethos of Sustainability Transitions literature, to measure the quality of policies designed for the *GSE*. These are: Green House Gas Emissions (GHGE) Levels, *Renewable Market Share*, *Wholesale Prices*, *Tariff Prices*, and *Unmet Demand Days*. The *GHGE Levels* and *Renewable Market Share* are environmental indicators, whereas *Wholesale Prices*, *Tariff Prices*, and *Unmet Demand Days* are economic indicators.

The duality of the performance indicators provides a holistic analysis of the impact policies make in the *GSE*. Policy actions could greatly benefit the environmental indicators, for example, making all renewable electricity generator costs subsidised by the government. However, this would come with great economic detriment to the tax paying population of Victoria to support this subsidy. Conversely, we could remove all non-renewable electricity, but Victoria's renewable generators may not be able to meet the total demand of the market, and thus electricity from foreign (interstate) markets would be required, which historically increases the *Wholesale Prices* of electricity [92]. These five performance indicators allow for a more rounded assessment of the impact of policy actions in the *GSE*, and introduce to this study the complexities of designing policies that balance both environmental and economic goals. Each performance indicator is equally important in the experimentation and analysis throughout this study.

GHGE Levels

GHGE Levels (tCO_2e) was identified as a crucial indicator for the performance of policies for the *GSE*, as the development of environmentally sustainable policies requires the reduction of *GHGE Levels*. The *GHGE Levels* in the *GSE* are primarily caused by the domestic electricity consumption, and the fuel sources, such as coal or wind, that is used to generate the electricity.

During a given bidding round, when an electricity generator makes a successful bid to supply electricity, the *GSE* records the *GHGE Levels* emitted from that generator, and does so for all successful bidding generators. In this study, we will be using the average annual *GHGE Levels* per household from 01/01/2022 to 12/31/2050 will be used as a performance indicator for this study, with the aim to minimise this average. Using a per household value ensures we compensate for population growth when comparing the emissions levels of different years.

Renewable Market Share

The *Renewable Market Share* (*%*) is required to ensure that any policies that are developed are promoting the transition to sustainable electricity sources. For each bidding round, the *GSE* records the percentage of electricity that was generated from renewable electricity sources, and outputs a yearly average share. The mean of the yearly averages from 01/01/2022 to 12/31/2050 will be used as a performance indicator for this study, with the goal to maximise this mean value.

Wholesale Prices

Wholesale Prices (*$/MWh*) are computed by the *GSE* for each bidding round, and are strongly dictated by the supply and demand requirements of the system. The average wholesale price for all bidding rounds from 01/01/2022 to 12/31/2050 will be used as a performance indicator, to assess the economic impacts of policy actions, with the aim to minimise the price.

Tariff Prices

Electricity generated in the NEM sell their electricity to retailers, these are companies that act as a medium between generators and consumers. These retailers then sell the electricity to consumers, such as households or businesses. When a consumer receives a bill from the retailer for their electricity, the cost of the bill includes the cost of the electricity they used, and additional charges caused by external electricity system components, such as administration, maintenance, and other retail related fees. These additional fees are a tariff added to the bill.

The *GSE* does not simulate all components that make up the tariffs, but uses the *Wholesale Prices* to estimate the *Tariff Prices* using historic relationships between the *Wholesale Prices* and *Tariff Prices*. Using *Tariff Prices* as a performance indicator provides an insight on how policies may economically impact consumers, to ensure that consumers are not severely negatively impacted in the pursuit of a more sustainable electricity system. The *GSE* outputs an average annual tariff value for each household, and the mean of these annual values from 01/01/2022 to 12/31/2050 will be used as the performance indicator value, with the goal to minimise *Tariff Prices*.

Unmet Demand Days

In the event that the supply of electricity from generators is not able to meet the demand, the electricity is required to be imported from generators in other Australian states. Importing electricity exposes the Victorian electricity system to external electricity systems whose environmental, or economic electricity generation practices may not align with Victoria's. If this supply and demand imbalance occurs for a given bidding round, the calendar day that bidding round occurred on is marked as a day when electricity demand was unmet. The average annual number of unmet demand days for the simulation dates of 01/01/2022 to 12/31/2050 will be used as a performance indicator, to measure the ability for Victorian generators to meet Victorian demand, and will aim to be minimised.

4.3.1 Optimisation Problem

The combination of the minimisation/maximisation of each of the performance indicators can be formally described as a multi-objective optimisation problem. Consider the space of possible policy actions for the *GSE*, A, and the number of times a new policy action is activated in each *GSE* simulation, $N = 7$, as there are 7 political new cycles in each simulation period. Suppose $P \in A^N$ is the sequence of policy actions to be activated at the beginning of each political cycle. Let $f_{ghge}(P)$, $f_{renew}(P)$, $f_{wholesale}(P)$, $f_{tariff}(P)$, and $f_{unmet}(P)$ be the values of the performance indicators after executing P in the *GSE*. The optimisation problem is formally written as:

$$\min_{P \in P^N} (f_{ghge}(P),\ -f_{renew}(P),\ f_{wholesale}(P),\ f_{tariff}(P),\ f_{unmet}(P))$$

4.4 Policy Actions

To desirably influence the performance indicators of the *GSE*, careful consideration and research was conducted to determine what policy actions should be applied to the model by the MOEAs and RL agents in chapters 7 and 8. Following the work of [26, 64, 66, 97], multiple publicly available government and independent documents were identified that presented potential policy actions that can be applied to Victorian generators in the NEM [105–113]. The number of policy actions to be used could not be too large for this study, as this could potentially extend the runtimes for the MOEA and RL algorithms [114, 115], making them infeasible to use for this study. A final list of 10 different policy actions are in table 4.1.

Table 4.1: Policy actions available to implement in the *GSE*.

Name	Description	Impact
Carbon Tax	Generators pay a tax per tonne of CO_2 emitted. Tax prices are based on $23.00\$/tCO_2e$ in 2012 (historical carbon tax price [116]), and annually adjusted according to inflation.	Increases the LCOE of generators at a level proportional to their CO_2 emissions.
Secondary Market	Opens a hypothetical additional market, designed by Rojas-Arevalo [92], that allows priorities renewable generators to participate in a NEM market process.	Promotes the use of renewable electricity generators, reducing their LCOE.
Emissions Based Merit-Order	Changes the merit-ordering of bids in the NEM to order bids by the CO_2 emissions each bid will cause.	Prioritises renewable electricity generators, but doesn't guarantee lowest cost.
Increase Learning Rate (+15%)	Increases learning rate parameter (see table A.1), representing investment into renewable technologies.	Decreases LCOE of renewable electricity generators.
Increase Technological Improvement Rate (+15%)	Increases technological improvement rate parameter (table A.1), representing an expansion of renewable generator infrastructure.	Increases potential revenue for renewable electricity generators.
Renewable Electricity Subsidy (10%)	Subsidises renewable generators by reducing their electricity base price by 10%.	Decreases LCOE of renewable electricity generators.
Reduce Highest 1% Emitting Generators' Capacity (5%)	Reduces the generation capacity of the top 1% CO_2 emitting generators by 5%.	Reduces the potential revenue of highest emitting generators, increasing their LCOE.
Reduce Highest Emitting Generator Capacity (10%)	Reduces the generation capacity of the single highest CO_2 emitting generator by 10%.	Reduces its potential revenue, thus increasing its LCOE.
Reduce Highest Emitting Generator Capacity (20%)	Reduces the generation capacity of the single highest CO_2 emitting generator by 10%.	Reduces its potential revenue, thus increasing its LCOE.
No Action	No policy action is applied to the market.	N/A.

4.5 BAU Scenario

Throughout this study, repeated reference is made to a business-as-usual (BAU) scenario, that was used by Rojas-Arevalo [92]. The BAU scenario is a set of values for each of the input uncertainties in the *GSE*, that influence the model's outputs to best reproduce historical data. The BAU scenario is effectively a baseline scenario, and is intended to represent the values of input parameters if there was no possible uncertainty in the *GSE*. The BAU scenario is integral to the validation framework in chapter 5, and the construction of the reward function in chapter 8. Values for the BAU scenario can be found in Appendix A.1.

Chapter 5

Model Preparation

Careful consideration was given to how to best integrate the *GSE* [92], which is written in the Java programming language [103], with the Python programming language based experimentation platforms to be used in this study, the EMA Workbench [88] and RLlib [100]. This chapter discusses the preparation and integration of the *GSE* with the experimental platforms of this study, as well as optimisations to reduce the runtime of the *GSE*.

5.1 Python GSE

Rojas-Arevalo [92] employed the Python module JPype[1], to bridge communication between the Java and Python runtime environments for their experimentation with the *GSE*. JPype is a Python module that provides access to Java programs from Python by interfacing with both runtime environments using a shared memory approach.

The interaction between Python and Java occurs via the Java Native Interface (JNI), an interface exposed by the Java Virtual Machine (JVM) that allows for cross-language development. The JNI was used as a messaging passing interface between the Java and Python runtime environments, to pass data on the runtime conditions of the *GSE*, or to instruct the *GSE* on what policy actions to activate. Using the JNI was a costly overhead, and a cause for concern for this study, as the runtimes for MOEAs and RL algorithms are non-trivial, and can require millions of runs of a given model [114, 115]. As both the EMA Workbench and RLlib are Python based, Rojas-Arevalo's model was re-implemented entirely in Python, removing the complication of using the Java programming language with Python platforms. The runtime durations from 10 runs of the *GSE* implemented in Java and Python using a 2.0 GHz Intel Xeon server are displayed in table 5.1.

Language	Mean (s)	STD (s)	Min (s)	Max (s)
Java	77.04	3.08	74.07	83.64
Python	134.06	3.85	128.89	139.92

Table 5.1: Java vs. Python *GSE* runtime statistics from 10 runs of each program.

[1] https://jpype.readthedocs.io/en/latest/

Unfortunately, the resulting Python model was on average 74% slower than the original Java model. Java's speed can be attributed to the Just-In-Time (JIT) compiler the JVM uses, which compiles the source code into machine code for the JVM to use before runtime [103]. Conversely, Python is an interpreted language, with the source code converted to machine code at runtime [87]. Despite the convenience of having all computational aspects of this study in the same programming language, the speed of the Python model is not viable to be used for MOEA and RL experimentation, and as such, the original Java model will be used for this study. Nonetheless, the Python model will be provided to Rojas-Arevalo, to hopefully benefit future studies that intend to build upon Rojas-Arevalo's work.

Although the original Java model was significantly faster than the Python model, the $77.04s$ seconds per simulation is still a cause for concern. Even by exploiting multiprocessor architectures to simultaneously run multiple instances of the *GSE*, this runtime was still infeasible for the experimentation required to answer the research aims of this study. In order to be able to answer the research questions with a thorough investigation, the runtime needed to be reduced. The runtime of the *GSE* was profiled to better understand the computational demands of the different components of the model. This was completed using the Java profiling tool, JProfiler[2]. The profiling results indicated a vast portion of the runtime was spent on the creation and ordering of the bids for the merit-order market. As the number of bids in a given round is equal to the number of generators (<200), the time to order the bids of each round was insignificant. Instead, the computational bottleneck was due to how often this ordering occurred during runtime. In essence, the compute time for completing many quick tasks (bid ordering) added up to a much longer runtime.

5.2 Alternative Bidding Windows

As outlined in section 4.2, generators submit a bid in the NEM every 30 minutes. For 48 bidding rounds per day, between the simulation dates from 01/01/2019 to 31/12/2050, there are approximately $900,000$ bidding rounds for the entire simulation. Interestingly, the true NEM merit-order market has bidding rounds every 5 minutes, prompting the reasoning that the dynamics of the 30-minute bidding rounds simulated in the *GSE* was an approximation, or smoothing of the dynamics of the true NEM bidding rounds. The notion of further smoothing the bidding rounds by extending the time between bidding rounds presented an opportunity to reduce the model's runtime by a considerable amount, as longer times between bidding rounds would diminish the computational requirements of creating and sorting the bids.

To further understand the impact of smoothing the bidding rounds, simulations were run using different time intervals between bidding rounds, and the runtime values for various *GSE* attributes (e.g. *Renewable Market Share*) were recorded, and compared to true historical data. A similar historical validation was completed by Rojas-Arevalo [92], which will be used as a guide for the evaluation of the data for this investigation.

[2] https://www.ej-technologies.com/products/jprofiler/overview.html

Eight new time intervals between bidding rounds (bidding windows) were devised, with the condition that each calendar day contained a discrete number of bidding windows. These were 1, 2, 3, 4, 6, 8, 12, and 24 hours. The datasets used by the original model contained historical and forecast electricity demand, and solar capacity (amount of electricity a solar-powered generator could produce) for every 30-minute interval in the simulation, which aligned with the 30-minute bidding windows.

These datasets were also smoothed to align with the different bidding windows that were evaluated. Pseudocode for the algorithm to smooth the 30-minute forecast demand and solar capacity to a different bidding window size can be found in Appendix B.1. At a high level, the forecast demand for a bidding window of n hours was the summation of the forecast demand for the $2n$ 30-minute bidding windows it contained, and the solar capacity was the average solar capacity of the same $2n$ 30-minute windows.

In Rojas-Arevalo's model, electricity generators are only able to operate at a defined fraction of their true capacity. This defined fraction is known as their maximum capacity factor. For example, consider a generator with a production capacity of $30MW$, and a maximum capacity factor of 0.8. In a 30-minute bidding window ($0.5h$), the generator can submit a bid to supply $30MW \cdot 0.8 \cdot 0.5h = 12MWh$. This assumption was extended for the new bidding frequencies. For example, the same generator would be able to submit a bid of $30MW \cdot 0.8 \cdot 2h = 48MWh$ if the bidding window is 2 hours.

The validation framework compares the *GSE*'s outputs for annual total *GHGE Levels*, *Renewable Market Share*, *Wholesale Prices*, and *Tariff Prices*, with available historical data that accompanies Rojas-Arevalo's [92] source code. Validation of the model's ability to reproduce historical data is essential if the model is to be utilised as a foundation for exploring the use of policy actions in possible future scenarios. A successful validation was defined according to Rojas-Arevalo's original definition - "validation is considered successful when outputs follow the same trends over time than historical data and when statistical tests to quantify differences from historical and modelled data show the defined acceptable values" [92].

The years for historical validation time period occur between 1998 and 2020, a period that falls under the "Private Regime" of the electricity network, when the network operated under private ownership, as opposed to public ownership prior to 1998. Data is not present for all 23 years, but sufficient data is still present for each output to perform the validation.

$$MAE = \frac{1}{|S|}\sum_{i=0}^{|S|-1}|S_i - H_i| \qquad RMSE = \sqrt{\frac{1}{|S|}\sum_{i=0}^{|S|-1}(S_i - H_i)^2}$$

Two univariate statistical measures are used to assist in the validation process. The first, the Mean Absolute Error (MAE) [117] is used to measure the similarity between the simulated and historical data. Formally defined above, MAE is the expected value of the absolute differences between simulated (S) and historical (H) values. The second statistical value is the Root Mean Square Error (RMSE) [117], and was chosen to assist in identifying the impact of outliers in the results. Squaring outliers results in a disproportionate addition to the squared sum, relative to non-outlier values, thus having a greater impact on the final RMSE values. Similar to MAE, smaller RMSE values are desired, indicating smaller levels of residual difference amongst the datasets. An advantage to using MAE and RMSE in tandem is the relationship, MAE $\leq$ RMSE, such that as |MAE − RMSE| decreases, the impact of outliers decreases as well.

5.3 Historical Validation

The BAU scenario (section 4.5) was used to generate the validation data for the *GSE* for all evaluated bidding windows, as the BAU scenario is able to best reproduce historical values [92]. The results for *GHGE Levels* (total annual per household), *Renewable Market Share*, *Wholesale Prices*, and *Tariff Prices* values during the execution of the *GSE* are used for the validation framework, as true historical data for these values was publicly accessible [92]. Tabular data of the validation results discussed in this section can be found in Appendix B.2.

The primary aim of smoothing the bidding windows is to reduce the *GSE* execution runtime. Runtimes for each bidding window were determined by taking the average of 10 individual runs, and are presented in table 5.2. As expected, the runtime decreases, as the bidding window increases, due the reduced frequency of running the merit-order bidding cycle. These results will be invaluable when making the final decision of which bidding window to use for the remainder of this study.

	0.5hr	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
Runtime (s)	77.04	29.23	15.95	12.39	10.90	9.19	7.67	6.40	4.86

Table 5.2: Runtime duration of the *GSE* using different bidding windows.

5.3.1 GHGE Levels Validation

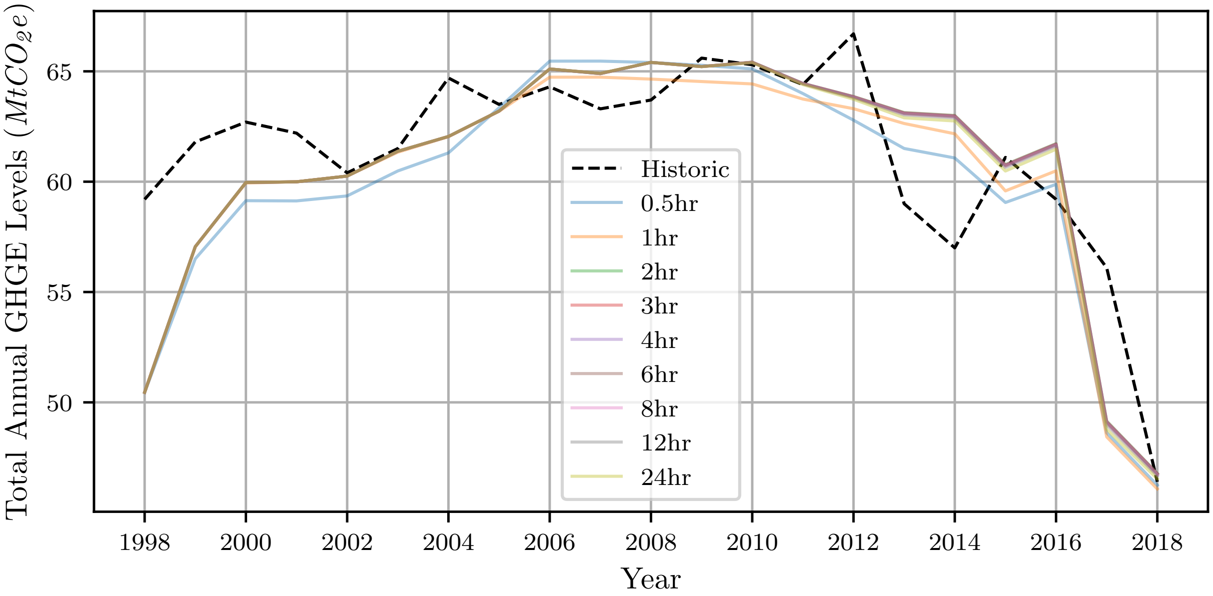


Figure 5.1: *GHGE Levels* from 1998 to 2018 using different bidding windows.

Figure 5.1 demonstrates the changes in *GHGE Levels* for the simulated data for each bidding window length, and the historical data. Visually, increasing the bidding window size appeared to have minimal impact on the *GHGE Levels*. The simulated results possess two periods of significant deviation from historical data that are both explainable. The first, occurs during the initial years of the simulation from 1998 to 2001, where the model's efficacy is reduced. When the simulation begins, all generators have a predefined LCOE according to their fuel type (e.g. coal), and require time to develop their LCOE that better reflects their own generation capabilities. The second period occurs at the beginning of 2012, when a carbon tax [116] was introduced in Australia, resulting in steep reductions of *GHGE Levels* from 66.7$MtCO_2e$ in 2012, to 59.0$MtCO_2e$ in 2013. The tax was repealed in 2014, resulting in the spike in historic data between 2014 and 2015. The *GSE* does not implement the carbon tax for the years it was active, thus explaining why the simulated *GHGE Levels* simulated were higher than the historic values over the carbon tax period.

($MtCO_2e$)	**0.5hr**	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
MAE	2.53	2.39	2.36	2.36	2.35	2.35	2.35	2.35	2.34
RMSE	3.44	3.37	3.39	3.39	3.39	3.39	3.39	3.39	3.39

Table 5.3: Bidding window validation statistics - *GHGE Levels*.

The statistical results (table 5.3) convey the minimal impact to *GHGE Levels* when the bidding windows are changed. Increasing the bidding window length produced more desirable values, where the largest window, 24-hours, possessed a MAE of $2.34 MtCO_2e$, and a RMSE of $3.39 MtCO_2e$. Compared to the original 30-minute bidding window (the poorest performing window), which had a MAE $2.53 MtCO_2e$ and RMSE of $3.44 MtCO_2e$, the absolute difference of MAE and RMSE was only minor between the best and worst performing windows, $0.19 MtCO_2e$ and $0.05 MtCO_2e$ respectively.

All bidding windows were considered to have acceptable MAE and RMSE values for the *GHGE Levels*. The RMSE values were all similar to the MAE values, indicating minimal presence of outliers in the data that would otherwise disproportionately skew the RMSE values. Therefore, we can conclude that all bidding windows lengths can appropriately reproduce historical data, to a degree close enough to use as a basis for future projection later in this study.

5.3.2 Renewable Market Share Validation

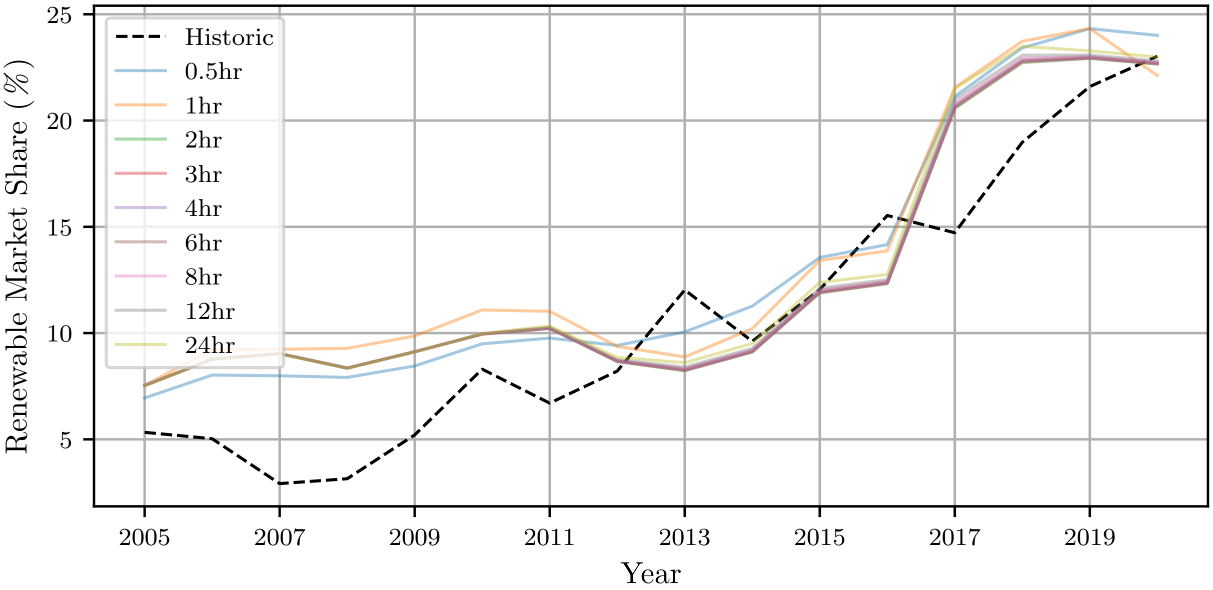


Figure 5.2: *Renewable Market Share* from 2005 to 2020 using different bidding windows.

Figure 5.2 displays the simulated annual *Renewable Market Share*, between 2005 and 2020, originally sourced from OpenNEM[3], a public platform for collecting NEM data. The simulated results reaffirmed the efficacy of the smoothed bidding windows, which closely followed historical values. The resulting MAE and RMSE values in table 5.4 increased as the bidding window increases, with the exception of the 1-hour window, which has the poorest performance, suggesting that smoothing the bidding windows comes at a cost of historical accuracy.

[3] https://opennem.org.au/energy/vic1/

(%)	**0.5hr**	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
MAE	2.76	3.36	2.86	2.86	2.86	2.86	2.87	2.88	2.92
RMSE	3.19	3.89	3.44	3.44	3.45	3.46	3.47	3.49	3.57

Table 5.4: Bidding window validation statistics - *Renewable Market Share*.

Comparing the two most different windows, the 30-minute window has a MAE and RMSE of 2.76% and 3.19% respectively, and the 24-hour has a MAE and RMSE of 2.92% and 3.57% respectively. The absolute difference between these two bidding windows is small, meaning that there was little deviation from historical values. For all smoothed windows, the MAE values are similar to their RMSE values, indicating the lack of outliers in the errors from the historical values. Accordingly, the impact of increasing the bidding window is minimal for *Renewable Market Share*, and with the small MAE and RMSE values present, all smoothed windows can be considered to appropriately replicate the historic data, and can be used for future exploration.

5.3.3 Wholesale Prices Validation

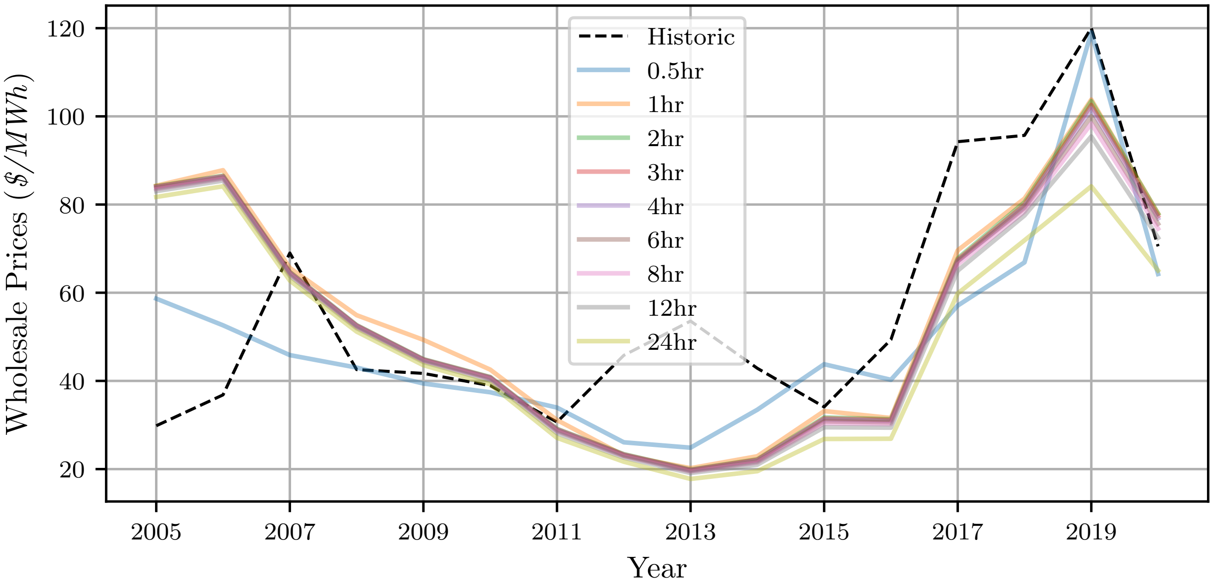


Figure 5.3: *Wholesale Prices* from 2005 to 2020 using different bidding windows.

In figure 5.3, the results for *Wholesale Prices* possessed multiple spikes throughout the validation period. Spikes can occur from multiple different factors, including supply/demand imbalances, macroeconomic and environmental factors (such as pro-longed extreme weather), and regulation policy. In the event a supply/demand imbalance occurs, the *GSE* assumes electricity is imported at a cost of 29% more than the last accepted bid price, where 29% is the historical average price difference of imported electricity prices [92]. All simulated results deviate more significantly from the historic values, compared to the deviation seen in the previous analyses of *GHGE Levels* and *Renewable Market Share*. As the *GSE*'s design makes broad assumptions on the macro and micro-economic behaviour of the NEM, the model lacks the precision to accurately reproduce historical data.

(*$/MWh*)	**0.5hr**	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
MAE	14.07	18.13	18.00	18.09	18.20	18.30	18.48	18.79	20.77
RMSE	18.31	24.07	24.00	24.06	24.14	24.21	24.37	24.67	26.33

Table 5.5: Bidding window validation statistics - *Wholesale Prices*.

Over the validation period, the MAE of all bidding windows except the 24-hour bidding window were between 17.00-18.00*$/MWh*, compared to the 14*$/MWh* in Rojas-Arevalo's model, this is an assuring minor deviation. The RMSE values all occurred around 24.00*$/MWh*, with the 24-hour bidding window being at 26.33*$/MWh*. The larger RMSE values suggest the presence of outliers in the results, which were likely caused by the prices over the carbon tax years, which can be clearly seen in figure 5.3, where the simulated prices did not follow the sharp rise and fall of prices.

The variation between bidding window sizes was minimal, indicating that the errors introduced by smoothing out the bidding windows resulted in an initial large deviation from the historical values, with the rate of change of this deviation decreasing as the window sizes were increased. Overall the change in MAE is still small, as shown in table 5.5, the MAE increased by only 6.70*$/MWh* from the 30-minute window to the worst performing 24-hour window. Figure 5.3 shows that increasing the bidding window size, caused the simulated *Wholesale Prices* to decrease relative to the other smoothed bidding windows, with the 24-hour bidding window prices having notably lower prices from 2017 onwards. This indicates that the omission of intra-day peaks, which are smoothed out entirely by a 24-hour bidding window, have a non-trivial influence on the *Wholesale Prices*. Whilst the impacts of smoothing the bidding windows are more pronounced for *Wholesale Prices* compared to *GHGE Levels* and *Renewable Market Share*, overall, the results indicate that any of the smoothed windows are still able to acceptably replicate the trends and to a lesser degree, historical values for *Wholesale Prices*.

5.3.4 Tariff Prices Validation

The *Tariff Prices* validation used historical data obtained in Rojas-Arevalo's dataset [92], for the years of 2001 to 2019. *Tariff Prices* were collected by Rojas-Arevalo from two different sources. The first data source is the St Vincent de Paul Society[4] (St. Vinnies), who have been recording annual average tariff prices since 2010. The second is a report submitted to the Australian Competition and Consumer Commission[5] (ACCC). Unfortunately, the lack of standardisation for recording *Tariff Prices* has resulted in a wide disparity between the data sources for years when they both have recordings [92].

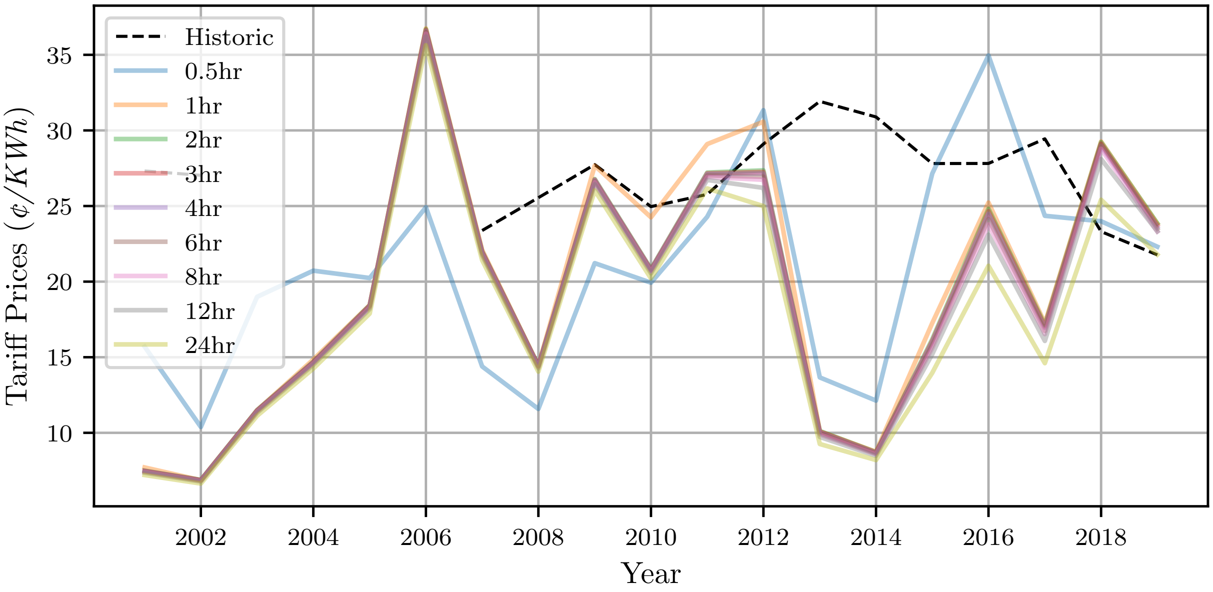


Figure 5.4: *Tariff Prices* from 2001 to 2019 using different bidding windows.

Figure 5.4 plots the simulated results against the average of the two historical *Tariff Prices* data sources. In the event *Tariff Prices* for a given year were not present in one of the data sources, the available *Tariff Prices* were used. Omitting years that did not contain data from both sources resulted in data only available for 2010 to 2017, where the simulated *Tariff Prices* highly deviated from historical *Tariff Prices* due to the *GSE* not capturing the economic impacts of the carbon tax. The bidding windows all maintained similar behaviour, except for a deviation in behaviour by the 1-hour bidding window, for the years of 2010, to 2012, where it exhibited larger *Tariff Prices* than all other simulations. For the years that contained historical data from both St. Vinnies and the ACCC, the simulated results were mostly between the two historical prices, except for 2013 to 2015, which is likely caused by the impact of the carbon tax on actual historical prices. Relative to the original 30-minute bidding window, the smoothed bidding windows still capture the same variations, increases and decreases in *Tariff Prices*.

[4] https://www.vinnies.org.au/page/Our_Impact/Incomes_Support_Cost_of_Living/Energy/VIC

[5] https://www.accc.gov.au/system/files/Victorian/Electricity/Distribution/Networks.pdf

(¢/*KWh*)	**0.5hr**	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
MAE	7.84	9.00	9.31	9.33	9.36	9.44	9.49	9.62	9.86
RMSE	10.09	12.12	12.24	12.27	12.29	12.35	12.40	12.51	12.88

Table 5.6: Bidding window validation statistics - *Tariff Prices*.

The MAE and RMSE results are displayed in table 5.6. The smoothed windows degrade in performance as the window lengths are increased, at an almost linear rate from the 1-hour window to the 24-hour window. However, the magnitude of the increase from 1-hour to 24-hours is not major for either of the MAE or RMSE (0.85¢/*KWh*and 0.76¢/*KWh*). After an initial degradation in simulated *Tariff Prices* by increasing the bidding window, further increasing the window does not have the same impact. Considering the relationship between the MAE and RMSE, it is expected the carbon tax years are the main contributors to their divergence.

Rojas-Arevalo noted sourcing historical data was challenging, and the lack of standardisation in *Tariff Prices* recording undermined the quality of the datasets. With regard to the smoothed bidding window results, despite some notable differences in raw values in the carbon tax years, all of the smoothed bidding window results can be considered to be appropriate for reproducing historical *Tariff Prices*. As stated, they were still able to effectively replicate the behaviour of Rojas-Arevalo's original model, with only a minimal skew.

5.4 Selection of Bidding Window

The results from the validation tests have demonstrated all evaluated bidding window sizes have an acceptable level of similarity to historical data for use as foundation for exploration into the future. The most notable phenomenon of smoothing the bidding windows is that the impact on the output indicators is relatively similar for all bidding windows. This suggests that a great deal of the variation in the outputs from *GSE* are caused by variations in demand and solar capacity that occur at the 30-minute bidding window interval.

From the analysis of the MAE and RMSE values, all bidding windows have been deemed to have acceptable levels of deviation from historic data. This study will seek to utilise the 24-hour bidding window period, as it possesses the fastest runtime of 4.86s on average from table 5.2. Using the fastest bidding window will be invaluable for the MOEA robust optimisation and RL training conducted in chapters 7 and 8. The experimentation to be conducted in both of these chapters are expected to require many hundreds of thousands to millions of runs of the simulation model [114, 115], to which is a serious challenge for the timeline of this study. In deep uncertainty literature, in addition to ensuring the runtime duration of computational models are short, it is common practice to also remove uncertain parameters from models if they have little influence on output values. Removing unimportant uncertain parameters reduces the complexity of designing policies using optimisation algorithms, such as Multi-Objective Evolutionary Algorithms.

Chapter 6

Augmenting Uncertainty

In deep uncertainty problems, if a parameter in the model being used has a trivial influence over the performance indicators, it should be removed from the set of uncertain parameters, reducing the uncertainty space of the model. To identify which of the 33 uncertain parameters of the *GSE* have the least influence, their influence must be quantifiable. Defining a parameter's influence on a performance indicator, can be considered the same as defining a performance indicator sensitivity to an uncertain parameter, which can be achieved by performing Sensitivity Analysis [118].

6.1 Sensitivity Analysis

Sensitivity Analysis (SA) is the process of identifying the contribution each uncertain input parameter has on the output values of a mathematical model [119]. SA is able to provide a comprehensive understanding of the relationship between uncertainties (uncertain parameter) and outputs of a model. For non-linear systems, this understanding becomes increasingly difficult as uncertain parameters can interact with each other, causing the total contribution of two interacting uncertain parameters to not equal the sum of their individual contributions [120]. The use of SA methods as a pre-processing step for models with uncertainty has increased over the last few decades, as the importance of minimising uncertainty where possible has become more apparent, and has in some cases contributed to identifying strategies to mitigate impacts of uncertainty on model outputs [120].

There are two types of SA, Local SA, and Global SA [121]. Local SA (LSA) is the targeted analysis of the influence of a single uncertainty on a model's outputs, by focusing on the relationship between the variance of the uncertainty, and output values, over multiple model evaluations. LSA is not able to handle non-linear systems, and as the *GSE* is a non-linear system [92], LSA techniques will not be considered for the SA of the *GSE*.

Global SA (GSA) provides a broader perspective of the influence of all uncertain input parameters, compared to the targeted local analysis performed by LSA. SA techniques are considered global when the values of all uncertain parameters are simultaneously varied throughout the analysis, and each uncertain parameter is varied over its full range. GSA techniques have been cited to be more suited to non-linear, real-world models, as they are able to examine the variation of multiple input parameters simultaneously [122]. GSA techniques will be used to assess the uncertain parameters in the *GSE*, to discern the level of influence the uncertain parameters have on the performance indicators, with the overall goal of fixing the value of parameters with little influence. There are multiple, well-known [119] methods for conducting GSA, including Sobol Indices [123], Fourier Amplitude Sensitivity Test [124], and Morris Screening method [125].

The Sobol Indices technique has been chosen for conducting the GSA of the *GSE*, due to the simplicity in interpreting its results, used in Rojas-Arevalo's original study, and widely adopted across the GSA literature [92, 119, 120, 126]. Sobol Indices [123] is a model independent, GSA technique, that is based on decomposing variance in the output performance indicators. Sobol Indices attribute the decomposed variances to input parameters in the model, thereby quantifying their influence. For this chapter, the required understanding of Sobol Indices is only the S1 and ST values, which quantify a given uncertain parameters contribution to the variance of output values. S1 values are calculated by evaluating the change in output values when only that given uncertain parameter is changed, whereas to calculate ST values, all uncertain parameters are changed when generating ST calculation data. By changing all uncertain parameters, any influence one parameter has over another parameter can be exposed when calculating ST. A detailed description of Sobol Indices can be found in Appendix C.2.

6.1.1 Experimental Setup

Computing Sobol Indices is supported by the EMA Workbench, which relies on an additional Python library, SALib [127], an open source library designed for performing SA. The sample size required for computing Sobol Indices are dependent on two main aspects of the model [121], the model's complexity, and the number of uncertain parameters. A set of 2100 scenarios with different combinations of the uncertain parameters will be sampled using Latin Hypercube Sampling (LHS) [128] (see Appendix C.1 for a discussion on LHS). This set size is double the minimum number of 1050 scenarios suggested by Gan *et al.* [119] for computing Sobol Indices, and was also the same number of scenarios used in Rojas-Arevalo's Sobol Indices analysis of the *GSE* [92].

The number of *GSE* evaluations (simulations) the EMA Workbench will run to compute the Sobol Indices follows a formula developed by Saltelli [129], which requires $C = n \cdot (2u + 2)$ evaluations, for a scenario set of size n, and u uncertain parameters. For the *GSE*, $C = 2100 \cdot (2(33) + 2) = 142,800$ evaluations were performed. For each year in the simulation, the first order Sobol Index (S1), the total order Sobol Index (ST) were recorded. These results were generated using a 44 core $2.0 GHz$ Intel Xeon server with multiprocessing, and took 3.12 hours to compute.

6.1.2 Discussion

Tables displaying the results of the Sobol Indices sensitivity analysis are presented in Appendix C.3 for each of the performance indicators. The median and maximum values for the recorded S1 and ST values for each uncertain parameter are recorded in each table, and are ordered by their median S1 value. The purpose of computing the S1 and ST values is to have quantitative aids for generating a list of the uncertain parameters, ordered by their importance to the performance indicators, that will be used by the factor fixing algorithm in section 6.2. Only the S1 index will be considered for the ordering process. The ST value is omitted as it considers the interaction of a given uncertain parameter with all other uncertain parameters, which will have less utility once some of the uncertain parameters are removed by the factor fixing algorithm in section 6.2.

By using the median S1 values, the ordering is less impacted by outlier years, where particular uncertain values may have a significantly larger influence on the performance indicators than other years. An example of this phenomenon occurs for the *generatorRetirement* uncertainty and the *GHGE Levels* performance indicator, which has a maximum S1 value that is far greater than its median. To combine the S1 values of all the performance indicators to a single, global list, the maximum median S1 value for each uncertain parameter was used for ordering. Other methods of combining the S1 values such as taking the mean or median values were considered, however, the maximum value was chosen as it ensured that any parameter that had a strong influence on at least one performance indicator was able to reflect this influence in the ordering, rather than potentially not representing this strong influence if the mean or median were used. The full list of 33 uncertain parameters ordered by their maximum median S1 value can be found in Appendix C.4.

As an additional filtering of the S1 values, the final list was compiled using a minimum cut-off S1 value of 0.01. Typically in Sobol Index literature, an arbitrary value of 0.05 is considered the minimum bound for significance [121], therefore using 0.01 ensures a conservative removal of uncertain parameters. An added benefit of removing uncertain parameters is the reduction in computational demand for the factor fixing algorithm in section 6.2, which has a linear relationship between the number of uncertainties and model evaluations. After filtering, only the top 16 most important uncertain parameters were kept for the *GSE*, and could be further reduced using the factor fixing algorithm.

6.2 Identifying Important Uncertainties

Using the ordered list of 16 uncertain parameters, the next step was to determine which uncertain parameters should be kept and removed (fixed) for later experimentation of the *GSE*. The factor fixing algorithm [130] iteratively evaluates the results of fixing the values for different ordered subsets of ordered uncertain parameters. Uncertain parameters that were fixed used their default values from the BAU scenario.

First, a set of scenarios S was sampled using Latin Hypercube Sampling (LHS) [128], and the output performance indicators P were recorded from running each scenario in S. The same set S was evaluated again, but only the n most important uncertain parameters used their sampled values, and the rest were fixed to their BAU values. The output values O_n were recorded, and the Pearson correlation coefficient was computed between P and P_n. This process was repeated for $n = 1..16$, where P_{16} had a correlation coefficient with P of 1. Detailed pseudocode for the algorithm can be found in Appendix C.5. Using a scenario set S, where $|S| = 2100$ and the 16 ordered uncertain parameters created in section 6.1.2, this algorithm required $C = 2100 + 16 * 2100 = 35,700$ model evaluations.

The results for each of the five performance indicators are plotted below in figure 6.1. The vertical axis is the Pearson correlation coefficient value, and the horizontal axis represents that the top n uncertain parameters used their sampled values, whilst the rest were fixed to their BAU values.

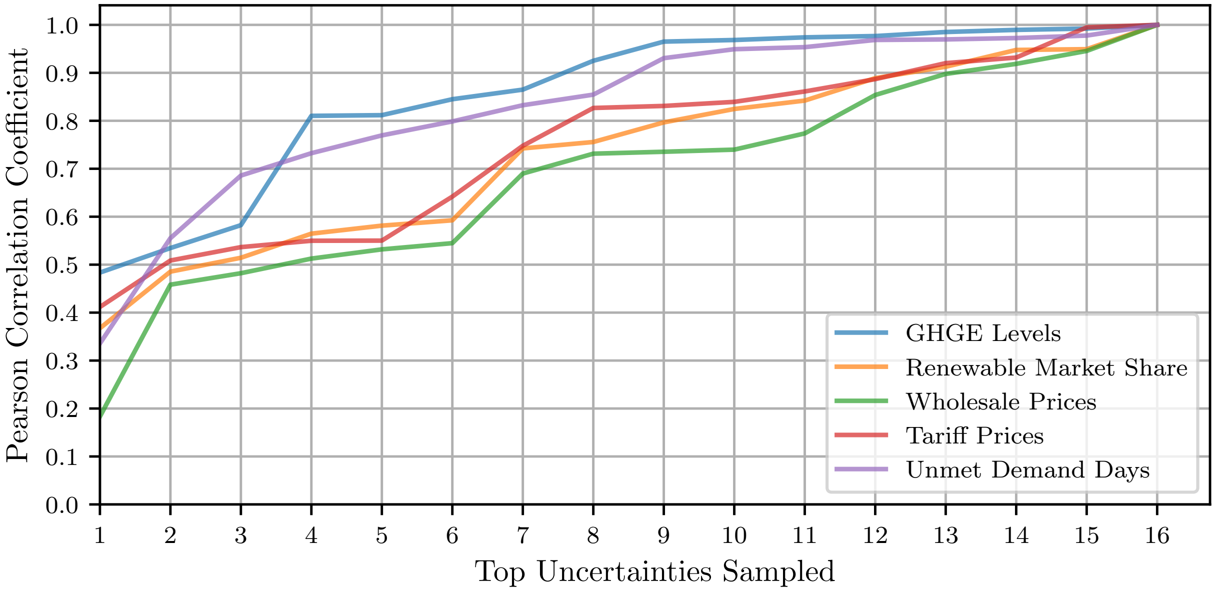


Figure 6.1: Pearson correlations per performance indicator when sampling up to the 16 most important uncertain parameters. Tabular data can be found in Appendix C.4.

The results in figure 6.1 demonstrate the effects of sampling the uncertain parameters in the *GSE*. By sampling only the most important parameter, *domesticConsumptionPercentage*, all the performance indicators have a correlation value of less than 0.5. This behaviour reaffirms the challenges of reducing uncertainty for non-linear models with deep uncertainty, and particularly for models with multiple performance indicators. Two notable spikes appear in figure 6.1. For the *GHGE Levels*, which has a significant increase of 0.23 once the fourth uncertain parameter, *priceChangePercentageBrownCoal*, is sampled. Price variations in brown coal are expected to have substantial impacts on *GHGE Levels*, as lower or higher prices would change the potential revenue, and therefore the LCOE of brown coal generators, directly influencing their bidding power in the merit-order market.

The other spike was for *Wholesale Prices*, when the second parameter, *nameplateCapacityChangeBrownCoal*, is sampled. Brown coal generators tend to be the amongst the largest and cheapest suppliers of electricity in the *GSE*. Changing a brown coal generator's electricity capacity would also influence their potential revenue and LCOE, again changing their bidding power in the merit-order. If the capacity is increased for brown coal generators, the ratio between brown coal generator capacity, and non-brown coal generator capacity would also increase, and the market share of brown coal generators is likely to increase as a result. This would result in more expensive generators not having bids accepted where they may in the BAU scenario, as the electricity demand has already been met by the now greater capacity brown coal generators. In contrast, a reduction in brown coal generator capacity may result in more expensive generators having successful bids to meet demand, allowing for more expensive generators to have successful bids.

6.2.1 Discussion

The evaluation has been conducted using figure 6.1 to assist in the process of choosing the number of uncertainties to use for later experimentation of the *GSE*. The Pearson correlation coefficient results in figure 6.1 were partially unexpected, as it was assumed that each of the performance indicators would have a substantial increase in correlation values during the addition of the first few most important uncertain parameters, and then the rate of change would drop steeply. Upon reflection, such expectations were unlikely to come to fruition in this experiment, due to the global ordering of the uncertain parameters. For example, the top five most important uncertain parameters for *GHGE Levels* is not guaranteed to be equivalent to the top five uncertain parameters in the global ordering.

The interaction between uncertain parameters and the impacts they have on the performance indicators may have also caused the more distributed increase. This behaviour would be able to be captured and understood by computing the Sobol Indices for the model with 16 uncertain parameters that were used in the factor fixing algorithm. Due to time constraints, this analysis is out of scope for this study, but would be strongly recommended for future work to strengthen this analysis. The final step of the factor fixing process is to determine the number of uncertain parameters that will be used for further experimentation. The trade-offs in making this decision is between the desired level of uncertainty (and implications on MOEA and RL convergence times), and the accuracy to reproduce the values of the true model that has all parameters sampled.

Figure 6.1 shows once the top 12 uncertainties have been added to the sampled set, the Pearson correlation coefficient for all performance indicators is greater than 0.8, indicating a considerable correlation to the original set [131]. Accordingly, the uncertain parameters for the *GSE* will be reduced to the first 12 parameters used in this analysis (table C.8). Using these 12 uncertain parameters minimises the levels of uncertainty in the *GSE*'s dynamics, whilst maintaining similar output values. This augmentation of the uncertainty for the *GSE* also hedges the risk of unfeasibly long runtimes for later experimentation in this study.

Chapter 7

Constructing the Metro Map

Optimisation algorithms are widely used for policy design for systems under deep uncertainty [17]. Generally optimisation algorithms are used to identify the best solution to a problem, but in deep uncertainty problems, such a solution rarely exists [132]. This chapter will explore the experimentation of Multi-Objective Evolutionary algorithms for the robust optimisation of policy pathways for the *GSE*, that will be used to construct a *Metro Map* that will guide an RL agent later in this study.

7.1 Multi-Objective Robust Optimisation

There are two main challenges for identifying the policy pathways to use for constructing a *Metro Map*: selecting pathways from a potentially infinite collection of possible pathways; and ensuring identified pathways are robust to the uncertainties in the system. A method to address both of these problems is to use Multi-Objective Evolutionary algorithms (MOEAs) for the robust optimisation of policy pathways with multiple performance indicators. MOEAs are well suited to optimising problems for non-linear spaces and high-dimensional action spaces [47, 133], and were used by Kwakkel *et al.* [57] to generate policy pathways for their original *Metro Map*. MOEAs solve optimisation problems by identifying solutions that are Pareto-optimal with respect to their objectives, where a solution x is Pareto-optimal if there does not exist a solution y, whose objective values are all better than the objective values attained by solution x [134]. To equate MOEA terms to concepts used in this study, a solution is a policy pathway, and the solution space is the space of all possible policy pathways. Similarly, MOEA objectives are the performance indicators stated in section 4.3.

To address the problem of ensuring optimised policy pathways are robust, we extend traditional MOEA optimisation to MOEA robust optimisation. Robustly optimising policy pathways can be considered as identifying pathways that produce desired outcomes for a variety of possible runtime conditions. In the *GSE*, different runtime conditions are defined by the set of input parameters, which are collectively called a scenario. As such, the robustness of a pathway is measured using the outcomes from a set of uncertain scenarios.

Algorithm 1 MOEA Robust Optimisation

```
 1: $S \leftarrow$ Set of randomly generated scenarios
 2: $R \leftarrow$ Initial random policy pathways
 3: $i \leftarrow 0$
 4: $P_i \leftarrow$ ParetoOptimal $(R, S)$          ▷ Performance based on multiple scenarios
 5: repeat
 6:     $C \leftarrow$ CrossOver $(P_i)$             ▷ Create child policy pathways
 7:     $M \leftarrow$ Mutate $(C)$                  ▷ Mutate children
 8:     $R \leftarrow$ Set of randomly generated policy pathways
 9:     $i \leftarrow i + 1$
10:     $P_i \leftarrow$ ParetoOptimal $(M \cup R, S)$
11: until termination condition met
12: return $P_i$                                    ▷ Robustly optimised policy pathways
```

MOEAs are inspired by biological evolution [135], closely following the evolutionary process. Pseudocode for how the MOEA will work in this study for robust optimisation is presented in algorithm 1. First, a set of random scenarios and pathways are generated (lines 1-2). The performance of the pathways (with respect to each performance indicator) is evaluated to identify Pareto-optimal pathways (line 4). The Pareto-optimal pathways are combined (cross-over, line 6) to produce new "child" pathways. The children are then slightly changed (mutated, line 7), and the mutated pathways are then called the population archive. Another set of random pathways are generated (line 8), and the Pareto-optimal pathways of the population archive and new random pathways are identified (line 10). This process is repeated until a termination condition is met. See section 7.2.3 and Appendix D.2 for the sizes of the sets S and R respectively.

7.1.1 ϵ-Non-Dominated Sorted Genetic Algorithm-II

The ϵ-Non-Dominated Sorted Genetic Algorithm-II (ϵ-NSGAII) [134] was chosen to facilitate the robust optimisation in this study, and is instantiated by algorithm 1. The ϵ-NSGAII algorithm is supported by the EMA Workbench, and is a successor to the Non-Dominated Sorted Genetic Algorithm-II (NSGA-II) [85], which was employed by Kwakkel *et al.* [57] to create their *Metro Map*. The key improvement of the ϵ-NSGAII algorithm over its predecessor is the reduction of the number of user-specified hyperparameters, making the ϵ-NSGAII a more accessible and simple algorithm to use. Specifically, it introduced adaptive runtime solutions to common hyperparameter specifications such as population sizing, and the inclusion of ϵ-dominance archiving [134].

The ϵ-NSGAII employs an iterative process where a population (set of solutions) is utilised for searching through the solution space, and population sizes are automatically adapted proportional to the number of Pareto-optimal solutions found. The ϵ-NSGAII keeps a subset of its Pareto-optimal solutions, by using a technique known as ϵ-dominance. The ϵ-dominance is a concept that allows the definition of precision for measuring the performance of identified solutions. When the ϵ-NSGAII is initialised, a grid is constructed over the objective space (figure 7.1), where the spacing between grid lines in each dimension is defined by a set of

"ϵ-precision" values. The grid is known as the ϵ-grid, and creates grid-blocks whose width in each dimension of the objective space is defined by the ϵ-precision values. Larger ϵ values result in a coarser ϵ-grid, and smaller ϵ values create finer grids.

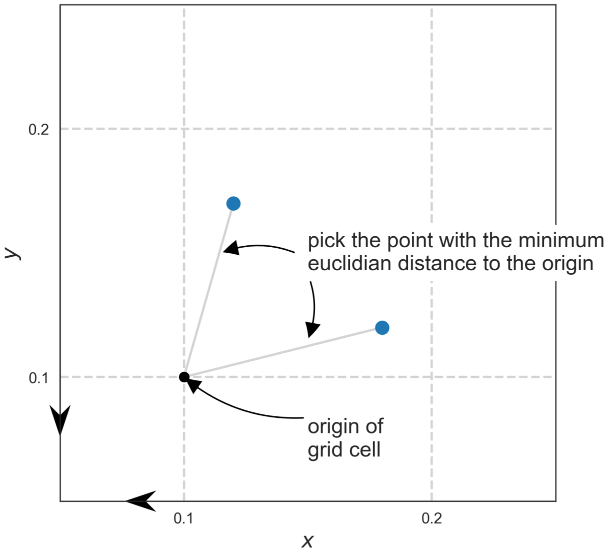

0.2
0.1
y
pick the point with the minimum
euclidian distance to the origin
origin of
grid cell
0.1
0.2
x

Figure 7.1: ϵ-dominance example for multiple solutions within a single ϵ-grid [88].

After each iteration of the ϵ-NSGAII, a grid-block in the ϵ-grid may contain the objective values from multiple solutions. This would occur if multiple solutions had similar objective values. If such an event occurs, the ϵ-NSGAII removes all solutions from that grid-block, except the one with the smallest euclidean distance to the origin (assuming that all dimensions in the objective space are to be minimised). See figure 7.1 for an example using a 2-dimensional objective. This reduction guides the ϵ-NSGAII to identify solutions with non-trivially different values across the objective space, promoting a more distributed search across the objective space. Since the removal of similar solutions by the ϵ-NSGAII is controlled by the ϵ-grid, the coarseness of the ϵ-grid directly impacts the population and therefore solutions kept after each iteration. Once all grid-blocks have been reduced to hold only one solution, the ϵ-NSGAII removes any solutions that are completely dominated by at least one other solution. This reduces the set of maintained solutions to only those that are Pareto-optimal, i.e., those that are ϵ-non-dominated. These solutions are collected to create the new population archive, whose role is to assist in the creation of the next generation.

Consider the population archive after iteration i, P_i (algorithm 1, line 7). To create the next generation of solutions to explore P_{i+1}, the ϵ-NSGAII uses a 25/75 creation rule, where 25% of the next generation are solutions from P_i, and 75% are randomly sampled from the solution space. The benefits of this creation rule are two-fold: it utilises ϵ-non-dominated solutions to guide the ϵ-NSGAII to re-use solutions who have previously proved their quality by ϵ-dominance; and randomly sampled solutions allow for exploration into unexplored areas of the solution space. Balancing the ratio of the create rule is very similar to the exploration/exploitation dilemma in Reinforcement Learning [16].

7.2 Experimental Setup

7.2.1 Defining Robustness

In MOEA robust optimisation, the robustness of a pathway is measured using the outcomes from a set of different uncertain scenarios. Typically, a robustness metric is used to map the values of a single performance indicator from multiple scenarios into a single robustness value. Thus, for the multi-objective problem of this study, after a given run of the *GSE* using a policy pathway, a robustness metric would be used five times to generate five robustness values for that policy pathway, one for each performance indicator.

The robustness values are then used to map a given policy pathway, to a point in the ϵ-NSGAII's objective space. The ϵ-NSGAII uses that point to conduct its ϵ-dominance checks between solutions, highlighting the importance of the robustness metric for how policy pathways are identified by the ϵ-NSGAII. An important consequence of the relationship between robustness metrics and the objective space is that the objective space is equivalent to the range of the robustness metric. The way in which robustness is measured for a pathway can be a precarious task, as two different robustness metrics can theoretically have widely varying opinions on the robustness of a given pathway. Multiple sources [45, 136, 137] suggested careful consideration is required when choosing a robustness metric, most notably, Kwakkel *et al.* who state further work should be completed to understand the influence of robustness metrics on MOEA robust optimisation [137].

Robustness metrics can be divided into three types: satisficing, regret, and statistical [137]. Given some predefined performance threshold, satisficing metrics seek to find pathways that maximise the number of scenarios that meet this threshold. An established example is domain criterion [138], which identifies the proportion of a pathway's results that fall within a defined area of the objective space. Regret metrics [139] are a comparative measure and aim to minimise the total regret of a pathway with respect to possible future scenarios. Regret has been defined as the difference between a pathway's performance in a given scenario, and the performance of the best performing pathway for that given scenario. A robust pathway in this context is one that minimises the maximum regret value across all evaluated scenarios. The third robustness family, statistical, investigates the statistical and distributional characteristics of the objective outputs from the set of scenarios. If the pathway's results are more skewed towards the desired objective values, the pathway is more robust.

The first two families of robustness metrics are not considered for this study. This study contends that predefining thresholds of satisficing metrics, and identifying best-performing pathways for regret comparisons to be contrary to the spirit of robustly optimising for a system under deep uncertainty, as this introduces new assumptions into the experimentation. As such, statistical metrics will be used to measure the robustness of policy pathways over multiple different scenarios.

A more intuitive way to think of robustness and robustness metrics, is to rationalise that the robustness of a pathway over a set of scenarios, is analogous to quantifying the overall performance of a pathway over those scenarios. A following conclusion is that robustness metrics are a way to encode different opinions on what makes a pathway's performance desirable. Below is an example that conveys this robustness/performance relationship for a simple, single-objective robust optimisation problem.

> Let P be the space of all pathways and S the space of all scenarios. Consider $p \in P$, whose performance is evaluated over the vector of scenarios $\vec{c} = \langle s_1, \ldots, s_n \rangle$, where $n \in \mathbb{Z}^+$, and $s_i \in S$, for $i = 1, \ldots, n$.
>
> Let the single performance indicator be $f(s, p) \in \mathbb{R}$, which is generated from using pathway p in scenario s. Collecting the outputs from evaluating p over $\vec{c}$, gives $\vec{o} = \langle f(s_i, p) \mid 1 \leq i \leq n,\ s_i \in \vec{c}\, \rangle$.
>
> Robustness metrics are functions mapping $\mathbb{R}^n \rightarrow \mathbb{R}$, and their task is to use the output vector $\vec{o}$ to define the robustness of pathway p, with respect to the scenario vector $\vec{c}$. Pathways are then ranked against each other by ordering pathways according to their robustness value, with the ordering dictated by whether $f(s, p)$ is desired to be maximised or minimised.
>
> Therefore, we can consider the process of quantifying the robustness of a pathway p for the scenarios in $\vec{c}$, is no different to quantifying the performance of p for the scenarios in $\vec{c}$. In summary, we can think of the terms robustness and performance for policy pathway evaluation as synonyms.

To extend this example to a multi-objective problem, consider the addition of another performance indicator $g(s, p) \in \mathbb{R}$. The robustness metric maps the results of $g(s, p)$ into a second robustness value. Therefore, each pathway has two distinct robustness values and rankings, one for each performance indicator generated by $f(s, p)$ and $g(s, p)$.

7.2.2 Robustness Metrics

Robustness metric taxonomies were reviewed with a specific focus to identify papers that discuss robust optimisation using MOEAs, and those that consider electricity and climate related scenarios [17, 45, 136, 137]. McPhail *et al.* contributed an excellent a list of robustness metrics ordered by their level of risk aversion (figure 7.2) [45]. This representation of robustness metrics allows the user to more clearly consider the consequences of their choices, and adjust their decisions according to their desired level of risk. It should be noted that the relative order of robustness metrics in this figure is subjective by its designers [45], and was included for illustrative purposes only.

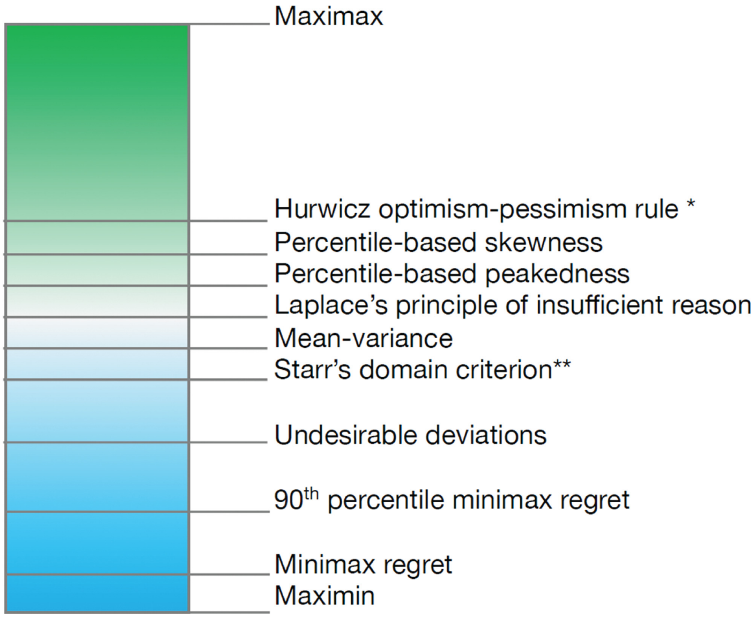

Maximax
Hurwicz optimism-pessimism rule *
Percentile-based skewness
Percentile-based peakedness
Laplace's principle of insufficient reason
Mean-variance
Starr's domain criterion**
Undesirable deviations
90th percentile minimax regret
Minimax regret
Maximin

Figure 7.2: McPhail *et al.*'s classification of robustness metrics in terms of relative level of risk aversion [45], where green denotes pro-risk, and blue is risk-averse.

Four different robustness metrics were chosen to assist the robust optimisation of policy pathways. Choosing multiple robustness metrics allows for the assessment of different risk attitudes for policy pathway design in the *GSE*, and a more comprehensive exploration of the policy pathway space by the ϵ-NSGAII. This is paramount for ensuring the *Metro Map* contains the most effective policy pathways for optimising the performance indicators. The robustness metrics are presented below, in order of increasing risk aversion.

Maximax

Maximax [140] is considered to be an optimistic, pro-risk metric. It uses the best recorded outcome as the value for robustness, making the assumption that the best possible outcome will realise. *Maximax* seeks to find a pathway $\boldsymbol{p^*}$ such that for n different scenarios:

$$p^* = \begin{cases} \underset{p\in P}{\operatorname{argmax}} \left(\max\left(f\left(s_1, p\right), \ldots, f\left(s_n, p\right)\right)\right), & \textit{maximisation} \\ \underset{p\in P}{\operatorname{argmin}} \left(\min\left(f\left(s_1, p\right), \ldots, f\left(s_n, p\right)\right)\right), & \textit{minimisation} \end{cases}$$

Laplace's Principle of Insufficient Reason

Laplace's Principle of Insufficient Reason (*LPIR*) [45] states that if no knowledge exists on the probabilities associated with different scenarios, the decision should be taken by assigning an equal probability to all scenarios. The robustness value is thereby calculated using the expected value from all scenarios. This metric is well-suited to deeply uncertain problems, as the absence of probability distributions for uncertain variables is one of the defining features of deep uncertainty [20].

$$p^* = \begin{cases} \underset{p\in P}{\operatorname{argmax}} \left(\frac{1}{n}\left(\sum_{j=1}^{n} f(s_j, p)\right)\right), & \textit{maximisation} \\ \underset{p\in P}{\operatorname{argmin}} \left(\frac{1}{n}\left(\sum_{j=1}^{n} f(s_j, p)\right)\right), & \textit{minimisation} \end{cases}$$

90th-Percentile Maximin

The 90th-Percentile Maximin (*90-P*) is a pessimistic metric that designates the robustness value to be the equal to the 90th worst percentile value from the evaluated scenario and was the robustness metric used in the first paper to develop a *Metro Map* computationally [57]. Readers should note this is not the 90th-Percentile Minimax Regret in figure 7.2. This study estimates the *90-P* robustness metric would be placed between the 90th Percentile Minimax Regret and Minimax Regret in figure 7.2. Consider $\left(f\left(s_1, p\right), \ldots, f\left(s_n, p\right)\right)_n$ to be the n-th percentile value for the output results of a pathway p.

$$p^* = \begin{cases} \underset{p\in P}{\operatorname{argmax}} \left(\left(f\left(s_1, p\right), \ldots, f\left(s_n, p\right)\right)_{10}\right), & \textit{maximisation} \\ \underset{p\in P}{\operatorname{argmin}} \left(\left(f\left(s_1, p\right), \ldots, f\left(s_n, p\right)\right)_{90}\right), & \textit{minimisation} \end{cases}$$

Maximin

The most pessimistic robustness metric, *Maximin* [140] uses the result from the worst recorded outcome for a pathway. This leads to risk-averse pathways, and cab be thought to identify a lower bound on performance.

$$p^* = \begin{cases} \underset{p \in P}{\operatorname{argmax}} \left(\min \left(f\left(s_1, p\right), \ldots, f\left(s_n, p\right) \right) \right), & \textit{maximisation} \\ \underset{p \in P}{\operatorname{argmin}} \left(\max \left(f\left(s_1, p\right), \ldots, f\left(s_n, p\right) \right) \right), & \textit{minimisation} \end{cases}$$

7.2.3 Scenarios for Robust Optimisation

A critical decision prior to using the ϵ-NSGAII for robust optimisation is determining the number of different scenarios to use to assess a policy pathway's robustness. This is vital for the quality of the policy pathways the ϵ-NSGAII will yield, as too few scenarios can lead to the robustness metric used by the ϵ-NSGAII to misrepresent the true robustness of the pathway. With respect to RL terminology, this can be considered as determining the size of the training data for the ϵ-NSGAII. This will be completed by performing a stability test [57, 101], which identifies the number of scenarios that should be collected for a robustness value to stabilise, that is, to not significantly change with the addition of more scenarios. Stability tests also assist in handling the trade-off between the runtime, and the quality of the robustness value. Stability tests are used to not only identify the minimum number of scenarios required, but to also assist with reducing the runtime in MOEA robust optimisation. The required runtime for the ϵ-NSGAII is directly proportional to the number of scenarios used for robust optimisation. A large number of scenarios guarantees the truest reflection of robustness, but incurs a non-trivial runtime cost. Typically, this trade-off is identified using a qualitative visual analysis [57, 101].

A set of 2000 scenarios was generated for the stability test, sampling over all 12 uncertain parameters using Latin Hypercube Sampling (LHS) [128]. Ten random policy pathways were also generated using LHS, and were evaluated for all 2000 scenarios. An example output of the stability test is show in figure 7.3, where the *LPIR* metric values of the *Renewable Market Share* indicator for 10 random policy pathways are plotted. The results for all robustness metrics and performance indicators can be found in Appendix D.1. In figure 7.3, the horizontal axis represents the number of scenarios used to calculate the robustness value (vertical axis). When the horizontal axis value is equal to i, the first i scenarios from the set of 2000 scenarios generated using LHS are used to calculate the robustness metric.

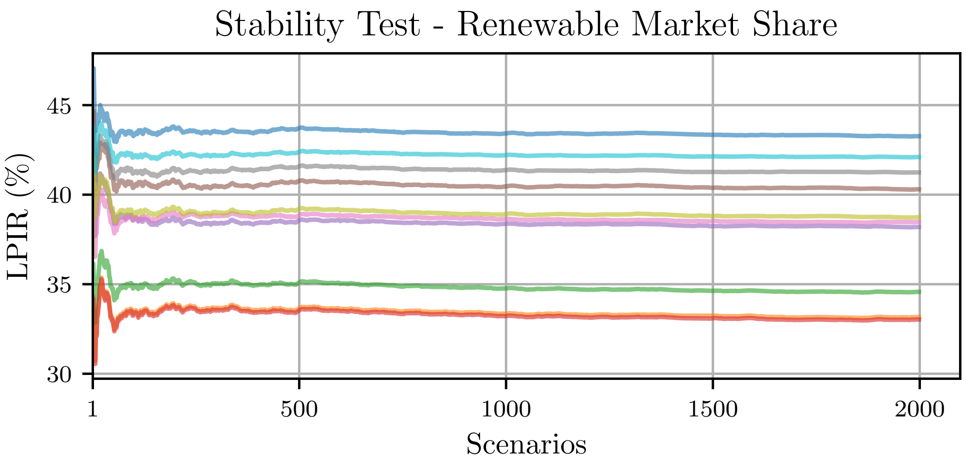


Figure 7.3: Stability test example - *LPIR* values of the *Renewable Market Share* indicator for 10 random policy pathways.

Between 500 and 1000 scenarios in the stability tests (Appendix D.1), the *LPIR* and *90-P* robustness metrics start to stabilise across each of the five performance indicators. Since the *Maximax* and *Maximin* are concerned with identifying the best and worst-case scenarios of a given policy pathway, the notion of stability is largely dependent on the ordering of the results, and theoretically may not stabilise until a vast proportion of the uncertainty space is explored. For example, if the scenario with the best outcome out of all scenarios is the first to be evaluated by the stability test, then the *Maximax* would appear to have immediately stabilised. Interestingly, the *90-P* which is a similar by design to the *Maximin*, appears to be much less sensitive to the addition of more scenarios than the *Maximin*.

Although the relationship between the runtime and number of scenarios for the ϵ-NSGAII is only linear, in the context of the expected runtime for the ϵ-NSGAII [137, 141], this is still of concern. Due to this, 500 scenarios will be used for the robust optimisation, as 500 scenarios reasonably stabilises for most metrics and performance indicators. Future work for this project would benefit from choosing a larger number of scenarios, to ensure with a greater degree of certainty that the robustness metrics have stabilised.

7.2.4 Sizing the ϵ-grid

The robustness metric used by the ϵ-NSGAII is highly influential on the behaviour and distribution of results in the objective space. For example, suppose that a given single-objective robustness metric, A, has a range of $[0, 10]$, and a second metric, $B = 2 * A$, which has a range of $[0, 20]$. If $\epsilon = 1$ for metric A, the ϵ-grid imposed would allow for at most 10 solutions to be identified, whereas for metric B, up to 20 solutions can be identified with the same width ϵ-grid. An ϵ-NSGAII experiment using metric B can be though to have a finer grid, and is able to conduct a more fine-grained exploration over the objective space compared to metric A. The conclusion to be drawn is that the robustness metric should guide the choice of ϵ values for sizing the ϵ-grid.

The ϵ values are generally determined via a trial-and-error process [88], requiring a balance between the runtime and solution quality. Smaller ϵ values increase both the number and quality of identified solutions [134], but more solutions require more model evaluations. For this study, none of the four chosen robustness metrics perform any form of dilation when mapping multiple performance indicator values to a single robustness value, either taking the mean or identifying a single scenario from the results. As a result, the size of the ϵ-grid is expected to have the same effect on all robustness metric experiments, as the coarseness of the ϵ-grid relative to the range of the robustness metrics is equivalent for all metrics.

Due to timing constraints, it was infeasible to tailor the ϵ values for each robustness metric. The ϵ values were determined via a trial-and-error process using the *LPIR* metric, and were considered to provide the best trade-off between runtime, quality, and number of pathways found. The ϵ values for each performance indicators for *GHGE Levels*, *Renewable Market Share*, *Wholesale Prices*, *Tariff Prices*, and *Unmet Demand Days*) were respectively $0.08tCO_2e$, 1%, $0.2\$/MWh$, $0.5¢/KWh$, and 1 *day*.

7.2.5 ϵ-NSGAII Termination

The MOEA literature contains various techniques on how to define when the solutions identified by an MOEA have (approximately) converged to signal termination. Three pervasive techniques were identified, Epsilon Progress [134], Hypervolume [142], and Epsilon Performance [134]. Epsilon progress uses the ϵ-grid imposed on the objective space as part of its measurement and is mainly used to signal the rate in which the objective space is being explored. Epsilon progress records when the ϵ-NSGAII has been able to identify a new solution that occupies a previously unoccupied grid-block in the ϵ-grid. Accordingly, Epsilon Progress is closely tied to the size of each generation, as larger generations have more capacity to explore new regions of the objective space.

Epsilon Progress is supported by the EMA Workbench, but only records the Epsilon Progress against the total number of pathways evaluated by the ϵ-NSGAII at that point in the runtime. Such results make it difficult to understand how the ϵ-NSGAII is progressing between generations, as the Epsilon Progress may appear to still be increasing considerably, when only a small fraction of the current population size is contributing to the Epsilon Progress. Additionally, Epsilon Progress does not provide any information into the change in the population, particularly, its quality and size, and therefore was not considered for this experiment. The other termination techniques, Hypervolume and Epsilon Performance, required either knowledge on the bounds of the performance indicators, or the creation of a reference solution respectively. As the bounds of some indicators (e.g. *Wholesale Prices*) are not known, and the notion of a reference solution introduces new assumptions to experimentation, these termination metrics were not considered.

Instead, this study will use the quality and size of policy pathway populations over multiple generations of the ϵ-NSGAII as a signal for termination. A similar approach was completed in [52], who explored electricity policy management in the European Union using MOEA robust optimisation. Below outlines the steps of the termination algorithm:

> Consider the population of policy pathways at generation i, P_i, where every $p \in P_i$ has five robustness values, one for each performance indicator. Record the mean robustness value for each performance indicator using all $p \in P_i$ and repeat this process for population P_{i-1}. Calculate the percentage improvement of each mean robustness value from P_{i-1} to P_i, resulting in five different percentage improvement values. If the maximum of these percentage improvements is less than $\Delta = 2\%$, then the ϵ-NSGAII has made little improvement from generation i to $i+1$. If for all $j = i-9..i$ (10 generations considered in total), the maximum percentage improvement value from P_{j-1} to P_j is less than Δ, the ϵ-NSGAII terminates.

By assessing the change over 10 generations, the ϵ-NSGAII has ample opportunity to test a sufficiently large number of randomly generated pathways (supported by the 25/75 creation rule), such that new pathways could be found that were non-trivially more robust and would influence the mean values of at least one robustness metric, as well as assisting the ϵ-NSGAII to move out of a local optimum within the objective space if needed.

This method to measure the termination of the ϵ-NSGAII maintains greater control over the change in quality of the pathways compared to Epsilon Progress, and is plotted against generation numbers, as opposed to the number of evaluated policy pathways, making it easier to understand the change in the solutions between generators. However, it has two deficiencies, the first being despite allowing up to 10 new generations to identify better solutions, it is possible that ϵ-NSGAII could terminate when it is stuck in a non-global optimum within the objective space. Secondly, the percentage change of robustness values can have different significance between robustness metrics. For example, a 10% change for *Maximax*'s indicates an average increase of 10% of the best case for all pathways, whereas for *Maximin* it would indicate a 10% increase in the worst-case. These deficiencies are not considered to have a notable impact on the underlying quality of the pathways' output, but are worth improving upon in future work.

7.3 MOEA Robust Optimisation Results

A set of 500 scenarios, as guided by the stability tests, was generated using LHS. These 500 scenarios were then used with the ϵ-NSGAII for multiple robust optimisations experiments, where each experiment used a different robustness metric from the four metrics outlined in section 7.2.2. These were run on a 44 core 2.0 GHz Intel Xeon server with multiprocessing, and collectively took 422.43 hours to complete (table 7.1). Even with multiprocessing to leverage the full computational power of the server, these runtimes reinforce the importance of model optimisations when experimenting with MOEAs [114, 115].

Robustness Metric	**Wall Time (hours)**
Maximax	75.27
LPIR	88.23
90-P	45.85
Maximin	213.08
Total	422.43

Table 7.1: ϵ-NSGAII robust optimisation runtime durations.

The hyperparameters for all ϵ-NSGAII runs can be found in Appendix D.2. Cross-over was completed using the simulated binary cross-over algorithm [143], and mutation by polynomial mutation [144]. The runtime behaviour of each experiment was recorded, such as the population size over time, as well as the performance indicator values for the robustly optimised pathways. Evaluating the runtime behaviour of each of the experiments will provide a deeper understanding of the influence robustness metrics have on the ϵ-NSGAII, particularly regarding the quality and number of policy pathways identified.

MOEAs are stochastic in their search through the solution space, introducing randomness into their algorithms. It is best-practice [137] to use the results of multiple runs of the MOEAs with different random seeds, so that the algorithm starts its first iteration with different sets of solutions. Initialising the exploration of the MOEA from different regions of the solution space mitigates the risk of the algorithm terminating whilst its solutions are stuck in a non-global optimum. The results in this experiment are only generated by a single run of the ϵ-NSGAII using each robustness metric due to timing constraints. Throughout the analysis, the results of each ϵ-NSGAII experiment will be referred to by the name of the robustness metric used in that experiment, for example, “the *LPIR* results...”, stands for the results of the ϵ-NSGAII that used the *LPIR* robustness metric.

7.3.1 Runtime Behaviour Results

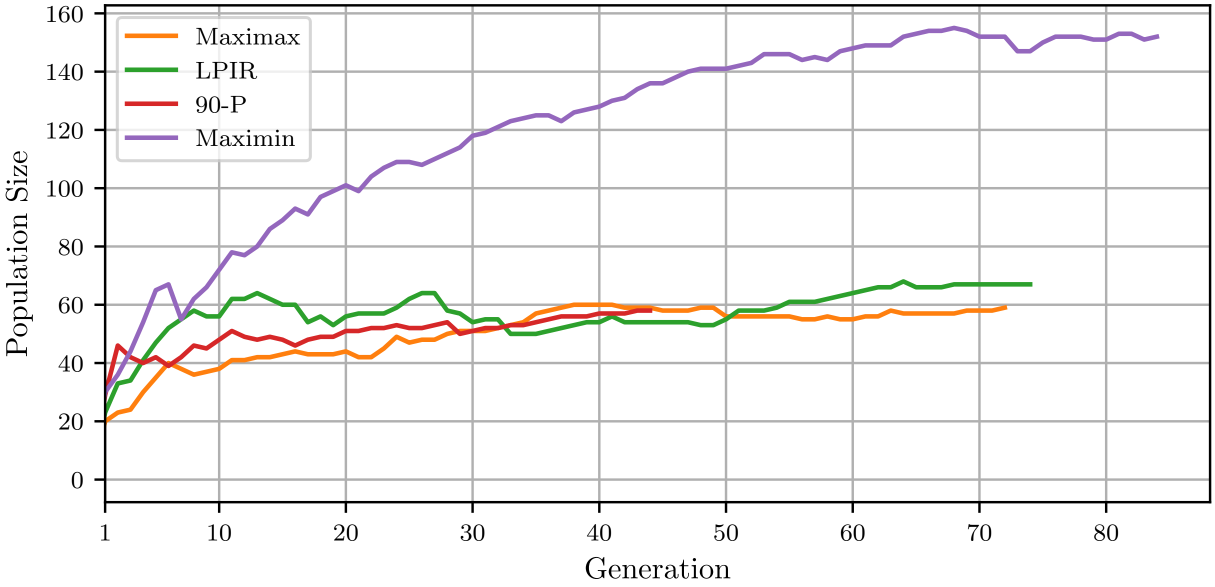


Figure 7.4: Population sizes throughout the ϵ-NSGAII robust optimisations for each robustness metric up to their respective termination generation.

The population size at each generation is plotted in figure 7.4. The *90-P*, *LPIR*, and *Maximax* metrics all maintained a similar population size throughout their execution. The upper outlier, the *Maximin* metric, contained a larger population sizes compared to the other metrics, with a maximum size of 155 policy pathways. This behaviour of the *Maximin* could reasonably be expected of the *Maximax* metric too, as it is effectively the reverse of *Maximin*, seeking the to maximise the best-case (outlier) outcome. These results indicate the objective space appears to be more densely populated in regions with less desirable behaviour, as the ϵ-NSGAII was able to identify so many solutions in the *Maximin* experiments. When approaching the more desired regions of the objective space, solutions become more sparse, making it challenging for the ϵ-NSGAII to identify as many pathways. From this, we draw that for the *GSE*, pathways tend to have a similar worst-case outcome, but best-case outcomes are more diverse, and difficult to identify.

Figure 7.5 shows the maximum percentage change between generations used for the termination algorithm (section 7.2.5), with all experiments reaching termination at the end of their plot (denoted by the dotted line). Both the *Maximax* and *LPIR* metrics exhibited very similar trends throughout their robust optimisation, especially over the course of the first 35 generations, averaging a maximum percentage change in metric values of 6.3% and 6.8% respectively.

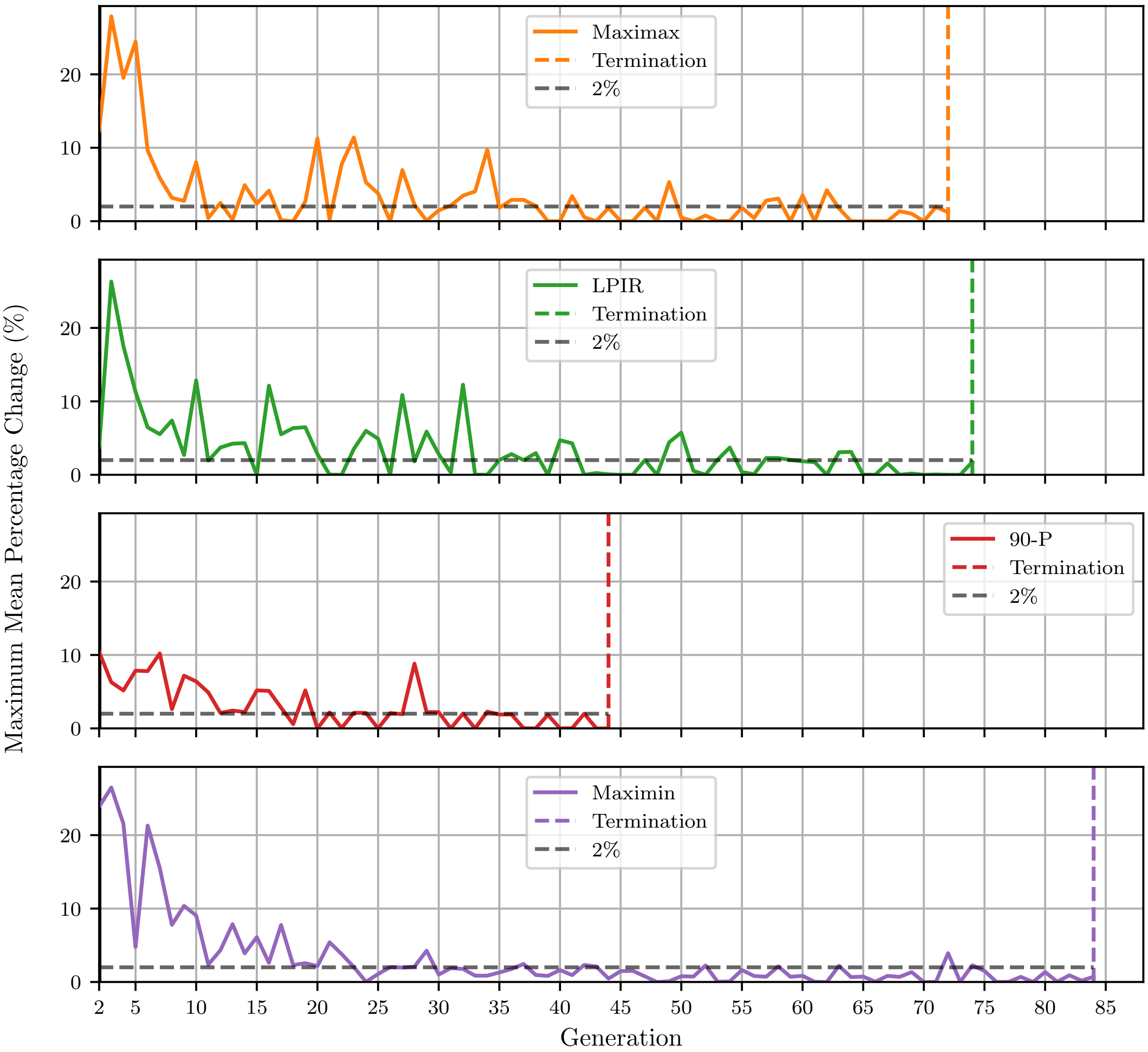


Figure 7.5: Percentage changes used by the ϵ-NSGAII termination algorithm for each robustness metric up to their respective termination.

The *Maximin* metric required slightly more generations to terminate (84), but has minor changes for a majority of its generations, very closely reaching termination multiple times after generation 20 for the following 64 generations. As the region of the objective space (worst-case) the *Maximin* explores is suspected to be densely populated, new pathways were frequently identified, whose robustness values were only minutely better on average by figure 7.5. Lastly, the *90-P* metric only required 44 generations to reach termination, significantly fewer than the other metrics. Given the ϵ-grid's are the same for all experiments, the variation in termination behaviour between the metrics was likely due to the robustness metrics themselves. The faster termination (w.r.t the number of generations before terminating) by using the *90-P* metric with the ϵ-NSGAII is a strong initial advantage of this robustness metric for the *GSE*.

7.3.2 Performance Indicator Results

Using the robustly optimised pathways from each metric, the performance indicator results were recorded for each policy pathway from the 500 scenarios used for the robust optimisation. The results were grouped according to which robustness metric they were optimised with, allowing for an investigation into the overall performance of using each robustness metric with the ϵ-NSGAII. Due to the final number of robustly optimised pathways varying between robustness metrics, the size of the grouped results (#pathways × #scenarios) varied as well. For example, the *90-P* metric dataset size is $48 \times 500 = 29,000$, and the *Maximin* metric dataset size is $152 \times 500 = 76,000$.

Robust optimisation evaluation should use a new set of scenarios, but the robust optimisation and subsequent creation of the *Metro Map* is all part of the training phase of the proposed DAPP/RL agent. In section 8.5 where the training phase is discussed, the same 500 scenarios used in this section are used as training data for the RL agents, thus we will continue to only use these same 500 scenarios for evaluation in this stage of the DAPP/RL agent's development. An additional baseline set of results is included for the following analysis. The data for the baseline was generated by running the 500 scenarios with no policy actions being taken at any point in the simulation. Completing this baseline analysis will assert whether the pathways identified by the ϵ-NSGAII were able to positively influence the performance indicators, compared to performing no action at all. Tables containing summary statistics of the results (e.g. mean, standard deviation) evaluated in this section can be found in Appendix D.3.

This analysis is guided by the words of Kwakkel *et al.* [102], who stated when evaluating the performance of policy pathways for systems under deep uncertainty, attention should be given to both the values and distribution of the performance indicators. Policies which are able to maintain a tighter distribution are viewed favourably, for their potential greater predictability and robustness to the uncertainties in the system.

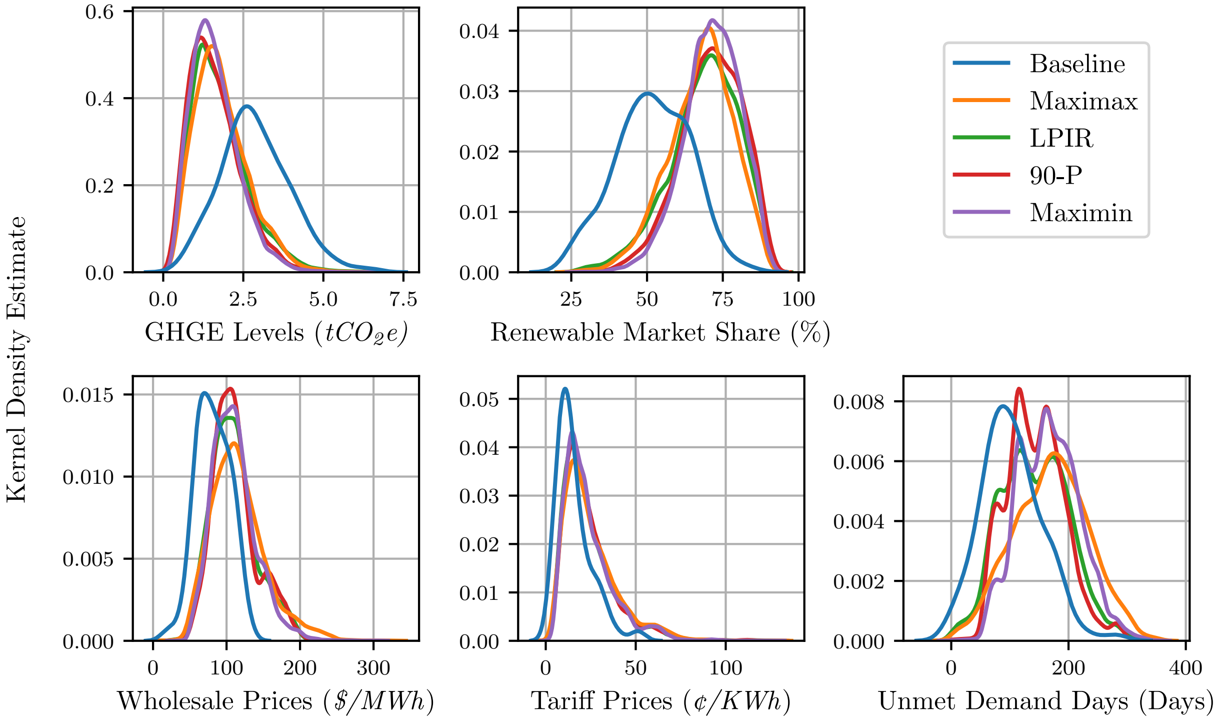


Figure 7.6: KDE plots of the results for all robustness metrics and performance indicators.

Kernel density estimates were fitted to all performance indicators results to provide a preliminary insight into the quality of the performance indicator results (figure 7.6). The environmental indicator results (*GHGE Levels* and *Renewable Market Share*) were greatly improved by all ϵ-NSGAII experiments compared to the baseline, but unexpectedly, all metrics produced very similar distributions for both environmental indicators. Considering how some robustness metrics have greatly differing ways of guiding the ϵ-NSGAII's search through the solution space (e.g. *Maximax* and *Maximin*), this study did not anticipate such similar behaviour. This distribution may have been caused by the policy pathways approaching an upper bound for overall environmental performance. Given the policy actions predominantly benefit the environmental indicators, it is likely the ϵ-NSGAII was able to identify policy pathways for each scenario that approached the upper bound of environmental performance. In contrast, the results between the robustness metrics for the economic performance indicators, *Wholesale Prices*, *Tariff Prices*, and *Unmet Demand Days*, exhibited less desirable results compared to the baseline. Reassuringly, their poorer performance was not as severe as the baseline for the environmental indicators. As the merit-order in the market is a cost minimising mechanism, the baseline was expected to have the best economic performance, as many of the policy actions (e.g. carbon tax) can have negative economic impacts. Regarding distribution, all metrics maintain a similar distribution for *Wholesale Prices* and *Tariff Prices*, but appear to spread slightly more for the *Unmet Demand Days* results. The results in figure 7.6 consolidate the key challenge of identifying policy pathways for the *GSE*, to find policy pathways that best balance the trade-off between optimising the environmental and economic performance indicators.

To convey the relative performance of the robustness metrics, the min-max normalised mean value of the performance indicators of all robustness metrics is plotted in figure 7.7. Caution should be taken when reviewing this figure, as it only displays relative differences. A first impression is a strong trade-off is present between the environmental and economic performance indicators for all results, as expected from figure 7.6. Such trends reaffirm the challenge to create policy pathways that balance the environmental and economic performance indicators.

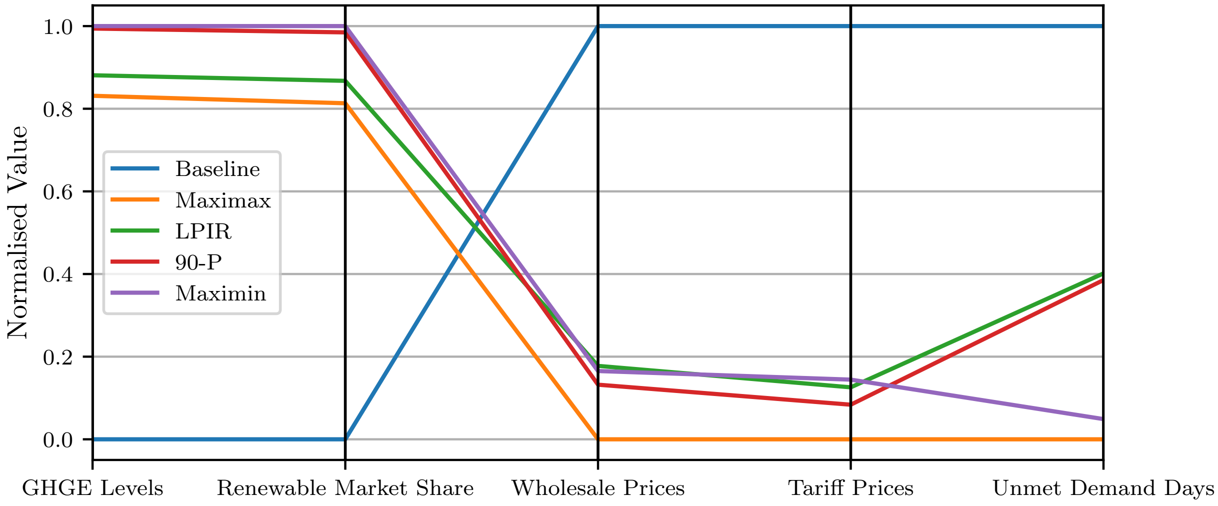


Figure 7.7: Min-max normalised mean results for all robustness metrics.

The pro-risk behaviour imbued by the *Maximax* metric did not pay off, exposing the ϵ-NSGAII to a suite of pathways who were the poorest performing on average for all the robustness metrics. In contrast, the risk-averse *Maximin* metric performed the best amongst the robustness metrics for most of the performance indicators, suggesting that given the choice between pro-risk and risk-aversion for the *GSE*, risk-aversion is a better strategy to adopt.

An interesting phenomenon occurred in the *Unmet Demand Days* results, where *Maximax* and *Maximin* split from *LPIR* and *90-P*. As was demonstrated in figure 7.6, the distribution of the *Unmet Demand Days* was diverse amongst the metrics, which is still present even when smoothing the results to their mean values. The key difference between the *Maximax/Maximin*, and the *LPIR/90-P* is the exposure their calculations have to the distribution of a policy pathway's results. *Maximax* and *Maximin* use the performance indicators from the best and worst-case scenarios, whereas the *LPIR/90-P* are implicitly exposed to the distribution of the results, as the mean value and 90$^{\text{th}}$ percentile are influenced by the distribution. In addition, both *LPIR/90-P* are by design less susceptible to outlier influences when calculating robustness. This study shows that identifying policy pathways that maximise performance for the *Unmet Demand Days* requires a more considered approach to the overall results of a pathway, taking into account both performance indicator values and distribution, as opposed to acting greedily or conservatively according to the values.

From only observing figures 7.6 and 7.7, the study finds the best performing robustness metric was the *90-P* metric, which presented desirable results across all performance indicators, and most pronounced for the *Unmet Demand Days*. The close second is the *LPIR* metric, who performed similarly across all performance indicators, but (relatively) less desirably for the environmental indicators. It is worth taking note of the risk aversion of the two best metrics are upper bounded by a risk-neutral metric (*LPIR*), and lower bounded by the risk-averse *90-P* metric. This reinforces the notion to maintain a risk-averse attitude for the *GSE*. Future work should evaluate a wider variety of robustness metrics, particularly those with a pro-risk level between the *Maximax* and *LPIR* to explore the performance of pro-risk policies for the *GSE*, and further investigate the concept of using risk-aversion as a bound for performance.

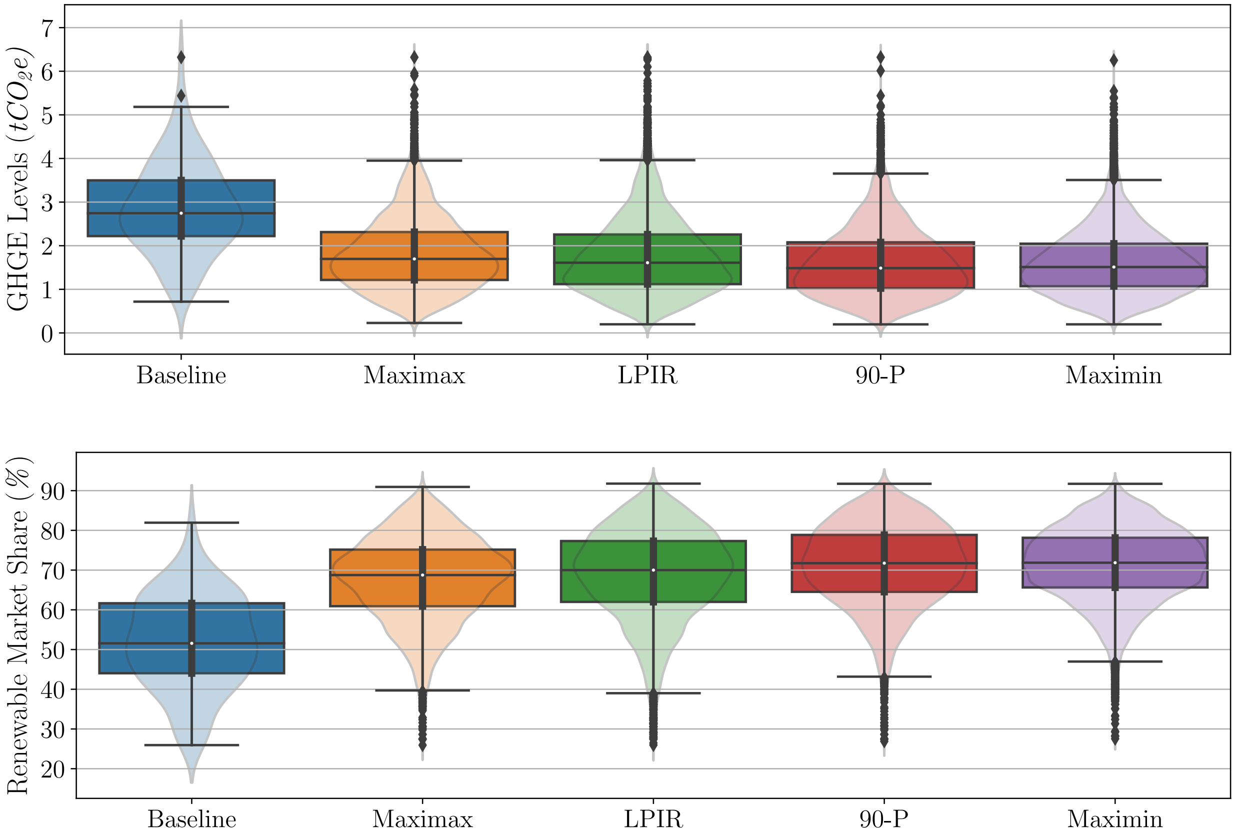


Figure 7.8: Robustness metric results for 500 scenarios - Environmental indicators.

Box plots were generated for each of the performance indicators to better visualise their quality. Figure 7.8 clearly demonstrates the capability of the ϵ-NSGAII to optimise policy pathways for the environmental performance indicators. The best result for *GHGE Levels* was a reduction of the mean baseline *GHGE Levels* value of $2.87tCO_2e$ to $1.61tCO_2e$, and the mean baseline *Renewable Market Share* was increased from 51.74% to 71.37% (both results were generated by the *Maximin* metric). Unfortunately, none of the results from the ϵ-NSGAII were able to notably improve the worst-case values for both environmental indicators, suggesting there are a set of unavoidable worst-case scenarios. However, a considerable improvement in the environmental performance indicators has occurred overall. An added benefit of the ϵ-NSGAII environmental results were its tighter distributions. Using the *Maximin* metric, the standard deviation was reduced from the baseline *GHGE Levels* value of $1.06tCO_2e$ to $0.74tCO_2e$, and the *Renewable Market Share* standard deviation was reduced from 11.88% to 9.29%. Although the changes in the distribution are not as significant as the changes of the raw values, the ϵ-NSGAII has identified policy pathways that can handle the deep uncertainty of the *GSE* and consistently ensure strong performance indicators. The improvements on the raw values and distribution of the policy pathways identified by the ϵ-NSGAII affirm that the ϵ-NSGAII has successfully designed environmentally beneficial policy pathways for this model, that hopefully result in an environmentally conscious RL agent that uses these policy pathways.

Shown in figure 7.9, the ϵ-NSGAII was not able to generate more desirable results for the economic performance indicators. Whilst unfortunate, this is an expectation of the policy pathways, as many of the policy actions have negative impacts on the economic indicators. The values for the *Tariff Prices* are the most displeasing, as the worst-case scenarios for each metric reach much less desirable prices than the baseline, which is likely due to the use of the more economically impacting policy actions, such as the merit-order change action. Across all economic indicators, the *Maximax* performed the poorest for all metrics, possessing the large mean values, range, and standard deviation (see Appendix D.3).

Notable behaviour arose from the economic indicators whereby all results had a similar lower bound on results (lower bound being desirable), suggesting the ϵ-NSGAII approached some bound for the best-case scenarios of its policy pathways. The *Maximin* is one exception, having the largest minimum value for *Wholesale Prices*, $32.22\$/MWh$, nearly double the (approximately) $17.00\$/MWh$ of all other minimum *Wholesale Prices*. Considering the design and purpose of the *Maximin* metric, the risk aversion may have influenced this result, as the metric does not consider the best-case performance of policy pathways, or any performance outside the worst-case, thus it is not unexpected for the *Maximin* to have poor best-case performance. Another notable result is again concerning the *Wholesale Prices*, and the impressive range of the *LPIR* and *90-P* metrics, both approximately $219.00\$/MWh$, far below the *Maximax* and *Maximin* ranges of $324.44\$/MWh$ and $306.92\$/MWh$ respectively. Both the *LPIR* and *90-P* metrics have been able to contain all results with a much tighter range, and were far less susceptible to outlier results compared to the other metrics. The range is an important statistical value for *Wholesale Prices*, due to the potentially devastating impacts of high *Wholesale Prices* can have for households and business the NEM supplies electricity with. Ensuring the range and maximum wholesale price value is as small as possible is essential for any policy pathway for the *GSE*.

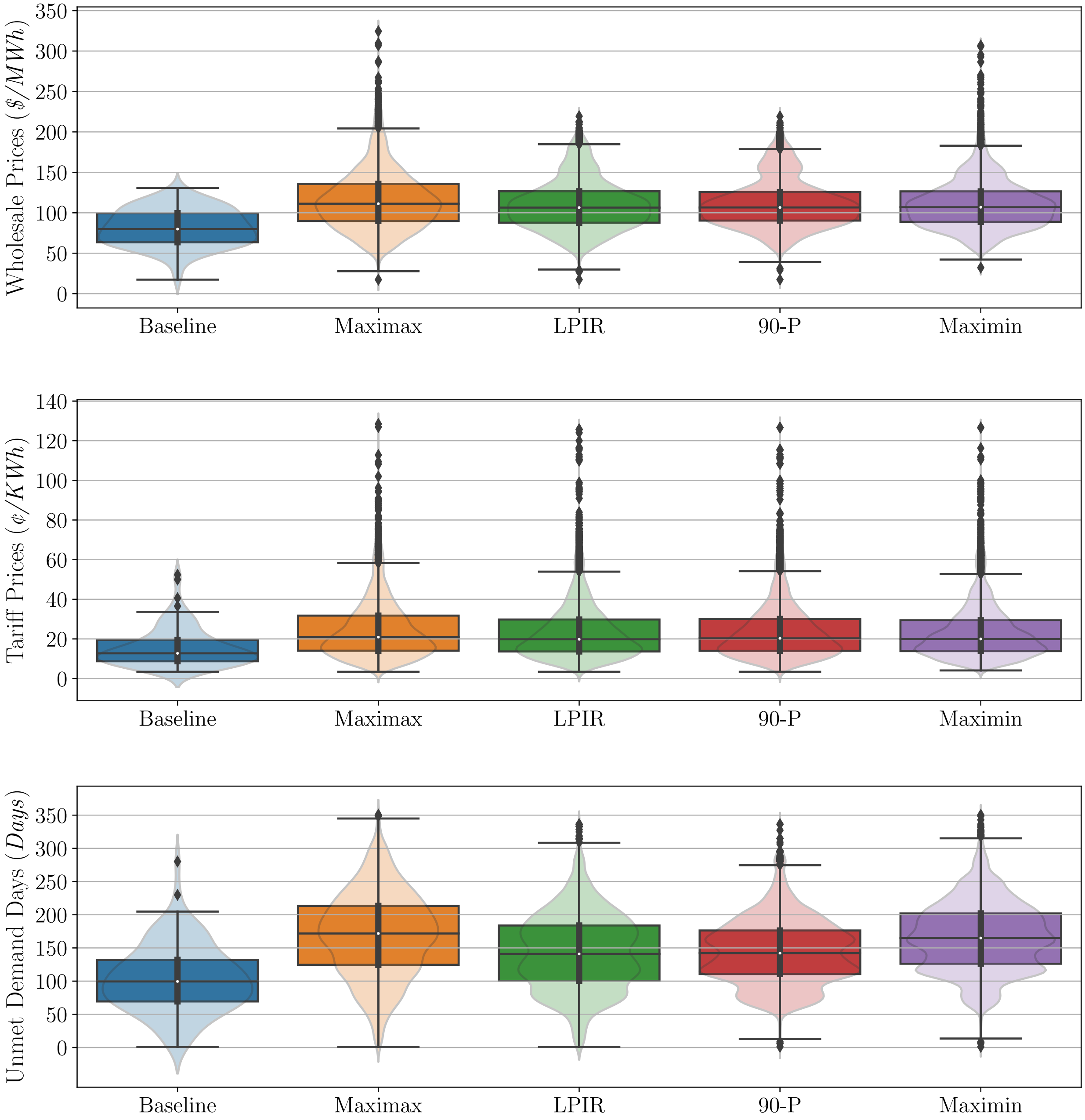


Figure 7.9: Robustness metric results for 500 scenarios - Economic indicators.

7.4 Discussion

Upon reflection, the robustness metric results relative to the baseline are not as poor as the analysis may suggest. Within each robustness metric, the results were generated by evaluating the performance of all policy pathways, against all scenarios. However, not every policy pathway is suited to every scenario, there is a need to be adaptive, to respond to the uncertainties and circumstances in each scenario and change policy actions accordingly. The aim of the DAPP/RL agent is to introduce this adaptive ability, to use the policy pathways to guide the agent, and provide the agent with a set of decisions in the form of the *Metro Map* that have already been robustly optimised using the ϵ-NSGAII. The importance of this analysis was to understand how different robustness metrics can influence the set of policy pathways the ϵ-NSGAII identifies. As the policy pathways will be used to construct the *Metro Map* (section 7.5), the choice of robustness metric will have significant implications for the rest of this study. Although there were slight variations in performance across the evaluated robustness metrics, there was no metric that significantly outperformed the others. This is an unfortunate outcome for this experiment, and lends itself to the statement that choice of robustness metric for MOEA robust optimisation does not majorly influence the collective performance of robustly optimised policy pathways for the *GSE*. Granted, this does not extend to other aspects the robustness metrics influence, such as number of policy pathways found or time to termination, as these were more distinguished. The timing constraints of this study do not allow the training and evaluation of four DAPP/RL agents, where each agent uses a *Metro Map* constructed using the policy pathways from one of the four robustness metrics. Accordingly, the policy pathways from a single robustness metric needs to be chosen to construct a *Metro Map* with. The number of policy pathways identified, the raw performance values, and the design of the robustness metric are key aspects to guide the decision. The number of policy pathways can have negative impacts on the construction of the *Metro Map*, too many pathways can result in a very large, complex *Metro Map*, that could be difficult for the agent to learn. Conversely, too few pathways can make the *Metro Map* too simple that combining an RL agent with the *Metro Map* has no benefit, as the problem could be reduced to a common RL multi-armed bandit problem [16].

The first robustness metric to be discarded is the *Maximax* metric. It maintained the poorest performance over the mean values (section 7.3.2), and its pro-risk nature is inappropriate for a real-world system such as the NEM. The next metric to discard is the *Maximin.* This metric contains a very large number of policy pathways (153), and contained a considerably large range for *Wholesale Prices* too. The final two metrics, *LPIR* and *90-P*, are almost indistinguishable across their statistical results (see Appendix D.3), and also have similar numbers of policy pathways, 72 and 53 respectively. However, for the purpose of sustainability transitions, particularly, the focus on environmental performance indicators, this study will seek to utilise the set of policy pathways set generated by the ϵ-NSGAII using the *90-P* robustness metric to construct a *Metro Map*. The *90-P* robustness metric performed marginally better for the environmental indicators, has a smaller number of policy pathways, and its tapered risk-aversion was deemed to be more appropriate for a policy design problem where a large population (Victoria) is vulnerable to the policy actions implemented.

7.5 Constructing the Metro Map

A difficult challenge of using MOEAs for robust optimisation is how to handle and utilise the multiple output solutions. Once a set of robustly optimised policy pathways are identified, human policy-makers must weigh up the pros and cons of each pathway and make a decision on which to implement. Such a decision is restrictive and, regrettably, loses the collective insight and knowledge the discarded policy pathways may have.

The *Metro Map* combines multiple policy pathways into a single graph, similar to a decision tree, which can increase the complexity of the decision faced by policy-makers. Instead of choosing a single policy pathway to follow and definitively knowing the rest of the policy pathway, the decision becomes which action to take, whilst considering the multiple possible policy pathways that can be followed after taking that action. Haasnoot *et al.* [54] argued this was a benefit of the *Metro Map*, stating that this allows policy-makers to commit to actions in the short-term, whilst providing a framework to guide future actions. Most importantly, plans can be made that can dynamically adapt as the future unfolds.

Once the set of robustly optimised policy pathways has been generated, constructing a *Metro Map* is a relatively trivial task. Pathways are combined by grouping together pathways who have the same "prefix" of policy actions. For example, if the first n policy actions in two policy pathways are identical, then those two pathways have the same "prefix" for the first n actions. These pathways can then be combined for the first n policy actions, and diverge in the *Metro Map* for the $(n + 1)$th action.

Figure 7.10 shows four different policy pathways, which are comprised of two different policy actions. Their prefixes are combined in figure 7.11, to construct the *Metro Map* (or decision tree). A black node is added to figure 7.11 as the root node, its purpose is to act as a starting point before the first decision is made.

Figure 7.10: Four simple policy pathways.

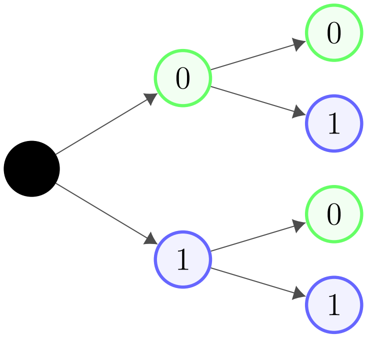
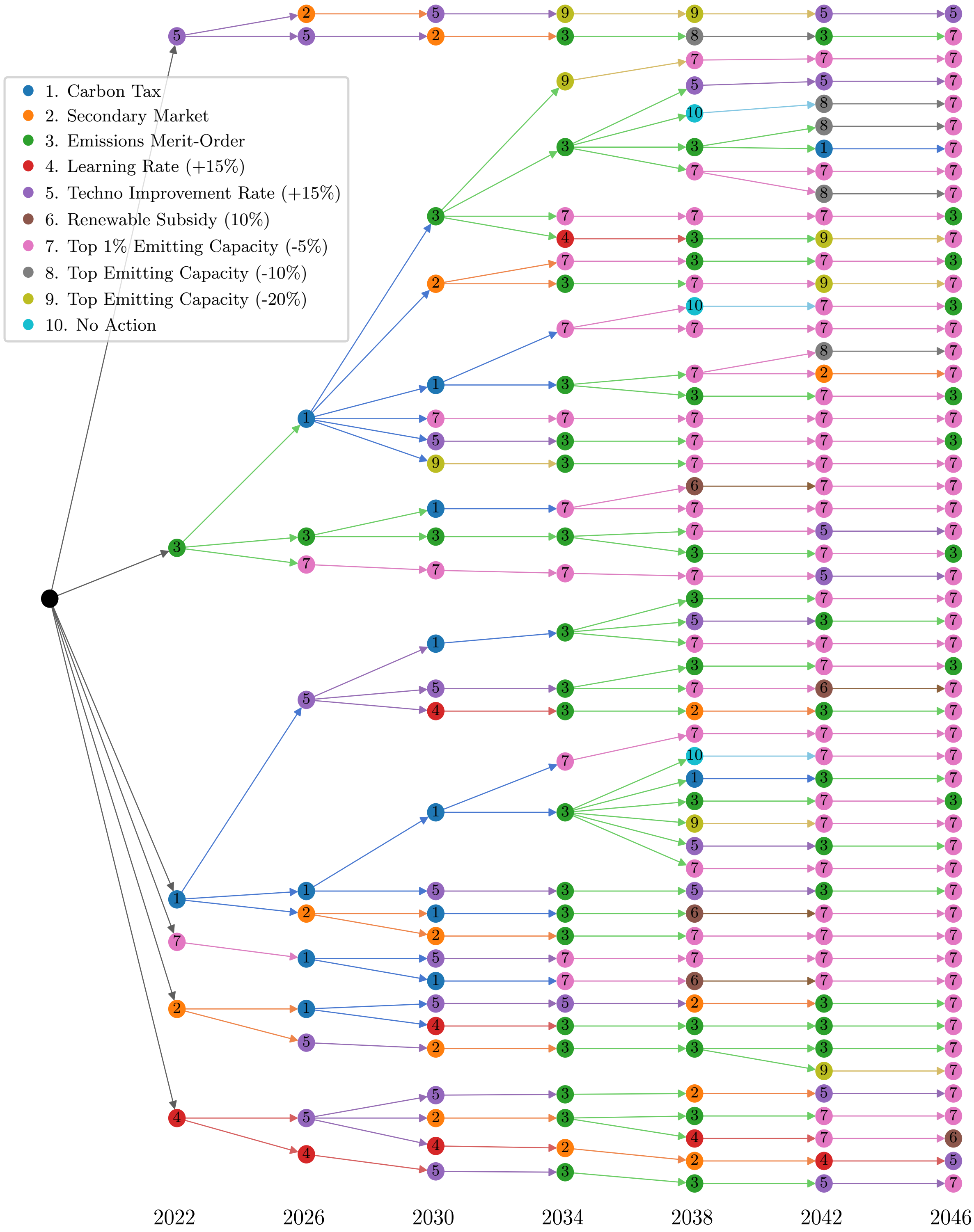


Figure 7.11: A simple *Metro Map* created using the policy pathways from figure 7.10.

Figure 7.12: *Metro Map* created using the policy pathways identified by the ϵ-NSGAII using the *90-P* robustness metric.


● 1. Carbon Tax
● 2. Secondary Market
● 3. Emissions Merit-Order
● 4. Learning Rate (+15%)
● 5. Techno Improvement Rate (+15%)
● 6. Renewable Subsidy (10%)
● 7. Top 1% Emitting Capacity (-5%)
● 8. Top Emitting Capacity (-10%)
● 9. Top Emitting Capacity (-20%)
● 10. No Action
2022
2026
2030
2034
2038
2042
2046

The 53 robustly optimised policy pathways generated by the ϵ-NSGAII using the *90-P* robustness metric were collected, and combined to make the *Metro Map* in figure 7.12. Traversal starts at the root black node, and every node is labelled and colour coded according to the policy action it denotes. Each level represents a different year in which policy actions are activated in the *GSE*, for example, the first level after the initial node represents the actions that can be taken in 2022, the following level 2026, and so on until the final level representing actions for 2046. The visual representation of the *Metro Map* in this study differs to the original design by Haasnoot *et al.* [54] due to the large number of policy pathways. Traditionally, *Metro Maps* are visually constructed to look like a public transportation map (hence the "metro"), but 53 pathways made the map difficult to present in this study. As a result, the visual representation of the *Metro Map* for this study is instead a tree structure.

There are two prevalent trends in the *Metro Map*, occurring in years 2034 and 2046 respectively. In 2034, a majority of the policy pathways activate the emissions merit-order action, and in 2046 almost all policy pathways activate the 5% reduction policy action. As the pathways were robustly optimised using 500 different scenarios, this would suggest that taking these actions in these years is often the best course of action, regardless of the uncertainties in the model.

Overall, the structure of the *Metro Map* is moderately complex, and has a sufficient branching factor such that applying RL to learn how to use the *Metro Map* will result in a better outcome, compared to randomly choosing any of the Pareto-optimal policy pathways within the *Metro Map*. In the second and third last levels, representing years 2038 and 2042, there is little to no opportunity to change policy pathways via a branching node. Branching would only occur if at least two policy pathways had identical actions up until these years, which is unlikely when there are 10 different policy actions to choose from. As a result, most of the RL agent's opportunity to adapt to scenario conditions will occur during the first four policy action timesteps, every four years from 2022 to 2034.

Chapter 8

Intelligent Policy-Makers

With the preparation of the *GSE*, and the construction of the *Metro Map* completed in chapters 4-7, this study can now begin the process of designing intelligent RL agents, whose purpose is to optimise the performance indicators of the *GSE* using the available policy actions. This chapter will discuss the preparation and training required to design the RL agents, and the considerations made to ensure the agents are able to handle the deep uncertainty of the *GSE*. The ability of the RL agents to optimise the performance indicators will be critically evaluated, and a discussion will be conducted on how the agents have addressed the research aims of this study.

8.1 Markov Decision Processes

Training intelligent agents using RL algorithms requires the interaction between the agent and its environment (*GSE*) to be mathematically formalised. In RL literature, it is common to use Markov Decision Processes (MDP) [16] as a framework to model this interaction. An MDP can represented as a tuple $\langle S, A, r, P, \gamma, \pi \rangle$,:

- S: State space that describes all possible configurations of the environment.
- A: Action space that describes all possible actions in the environment.
- $r(s, a)$: Reward function to determine reward given to agent after taking action $a \in A$ in state $s \in S$.
- P: Probability transition function to represent the environment's dynamics. $P_a(s'|s)$ denotes the probability of transitioning to state $s' \in S$, given the current state $s \in S$ and action $a \in A$ was taken.
- $\gamma \in [0, 1]$: A discount factor to assist in balancing immediate and future rewards.
- $\pi(s) \rightarrow a$: The policy (strategy) to decide what action $a \in A$ to take when the environment state is $s \in S$. Policies are generated when solving a MDP.

At each iteration of solving the MDP, the agent uses the current state of the environment s with its policy to decide which action $a = \pi(s)$ to take. The environment then transitions to state s' with a probability of $P_a(s'|s)$, and the agent is given a reward determined by $r = r(s, a)$. The following sections will outline the process taken to design the MDP to model the interaction between the RL agents and the *GSE* for this study.

8.2 State & Action Spaces

Designing the state space for non-trivial RL environments such as the *GSE* is a challenging task, as features to describe the state of the *GSE*, such as *Wholesale Prices*, are predominantly continuous values. When the state or action space is continuous, it is impossible to explore the product of these spaces to find the optimal policy for the MDP, described as the "curse of dimensionality" [16]. Another challenge is to design the state space such that it provides sufficient information for the RL agent to make informed predictions on the impacts and rewards of its actions [73].

Often in RL problems, the value of factors that should be in the state space to assist in action decisions are unknown. For example, having the true values of all 33 uncertain parameters of the *GSE* in the state space would greatly assist the RL agents, but would invalidate the investigation of how RL agents handle deep uncertainty. A common method to mitigate these unknown values is to employ Partially-Observable MDPs (POMDP), which attempt to model probability distributions of the unknown values to make more informed predictions of their true value.

For this study, we seek to understand how RL can assist in the policy-design process for deeply uncertain futures in the *GSE*. We find the best method to fulfilling the research aims of this study is to not fit probability distributions to attempt to learn the values of uncertainties in the model via mathematical techniques such as POMDPs. Rather, we aim to explore how RL agents can handle the deep uncertainty in the *GSE* that is caused by the 33 uncertain parameters. Further, fitting a probability distribution to uncertainties in the model would violate Maier *et al.*'s [20] principle of deep uncertainty that probability distributions of uncertain parameters are unknown or unable to be agreed upon.

To make informed decisions when interacting with the *GSE*, the RL agents require sufficient information [73] about different aspects of the *GSE* during runtime. Rojas-Arevalo [92] used a collection of 26 different values (Appendix A) that encompassed key aspects of the *GSE* to analyse the impacts of deep uncertainty via Exploratory Modelling and Analysis techniques. These values recorded yearly measurements (e.g. annual coal electricity production levels), and provided a comprehensive snapshot of the state of the *GSE* in each simulated year. These 26 values became the foundation of the state space for the RL agents. Some of these values were unrelated to the overall objectives of the RL agents, which could potentially mislead the RL agents to develop poor performing policies [16], and therefore some of the 26 values were removed, resulting in a final list of 20 values to describe the state of the *GSE* (Appendix A). The 20 values for each of the fours years were added to the state space, to capture the dynamics of the *GSE* in the four years between each policy action activation.

The final state space has a dimension of (4×20), 20 values for each of the four years, which are almost all continuous and capture the required information for RL agents to make informed decisions and predictions to optimise the performance indicators for the *GSE*. The action space for the RL agents is a mapping of the policy actions described in section 4.4 to a discrete integer set. As aforementioned, continuous state spaces introduce complexity problems for RL algorithms, and heavily influence which RL algorithms are applicable for this study.

8.3 Deep Q-Learning

The goal of the RL agents is to learn a policy π, for the MDP, which maximises the discounted sum of rewards in each episode (*GSE* simulation). The discounted sum, for an episode that is T timesteps long, is defined as $\sum_{t=0}^{T} \gamma^t r_t$. The discount factor, $\gamma \in [0,1]$, discounts the future rewards relative to the immediate reward. Discounting allows RL engineers to control the prioritisation the RL agent has between immediate and future rewards.

A classical approach to developing policies that maximise rewards for MDPs is the Q-Learning algorithm [145]. Central to Q-Learning is the Q-function, $Q^{\pi}(s,a)$, which measures the discounted sum of rewards which will be obtained when action a is taken in state s, and the policy π is followed until the episode ends. If the optimal Q-function, $Q^*(s,a)$, defines the maximum possible discounted sum of rewards that can be obtained by taking action a in state s, and continuing to follow the policy, then the optimal Q-function follows the Bellman optimality equation:

$$Q^*(s,a) = \mathbb{E}\left(r + \gamma \cdot \underset{a' \in A}{\operatorname{argmax}}\left(Q^*(s',a')\right)\right)$$

In simpler terms, the maximum reward return by taking action a in state s is the sum of the immediately returned reward, r, and the discounted sum of rewards generated by following the optimal policy until the episode terminates. The Q-Learning algorithm leverages the Bellman optimality by iteratively updating the Q-function according to:

$$Q_{i+1}(s,a) = \mathbb{E}\left(r + \gamma \cdot \underset{a' \in A}{\operatorname{argmax}}\left(Q_i(s',a')\right)\right)$$

It has been shown that as $i \to \infty$, $Q_i \to Q^*$, meaning that given sufficient iterations, the optimal policy will be found. By using the Q-function, RL agents are able to choose which action to take in a given state to maximise its returned reward. Rudimentary Q-learning algorithms store the Q-function in a tabular form, whose dimensions is determined by the size of the state and action spaces [16]. When at least one of the state or action spaces are continuous, tabular methods are no longer viable. A solution is to discretise the continuous space, but this can lead to a loss of information.

It has become common in RL literature to combine artificial neural-networks (ANN) to assist in combating this problem [13]. ANNs are able to mitigate the the complexities of continuous spaces by approximating different components of RL algorithms using neural-networks. ANNs also bring the benefit of their powerful feature extraction capabilities, which further assists in the state space design problem for RL [146].

This study will utilise the Deep Q-Learning (DQN) [147] algorithm to train RL agents to solve the MDP. DQN works by training a function approximator using an ANN with parameters θ, such that $Q(s, a; \theta) \approx Q^*(s, a)$. Using an ANN removes the use of tabular storage methods, and allowing DQN to operate with continuous state spaces. However, DQN finds the maximum of the action-value by enumerating all actions in the action space, limiting it to only being suitable for discrete action spaces. The DQN algorithm is well suited to the continuous state and discrete action spaces of this study, and has proven to be an effective RL algorithm for many other RL studies [148]. The DQN algorithm is known as a “model-free” algorithm, meaning it learns optimal policies by associating the best action with each state, without determining the underlying transition probabilities between states. As a result, this removes the need to formally model the probability function P of the MDP for this study.

8.4 Reward Function Design

Designing reward functions for RL problems is one of the most difficult and important tasks in RL studies. Sutton *et al.* [16] described the reward function as the method in which the user communicates its goal to the RL agent, where the challenge is to ensure the communication is clear enough to guide the agent towards that goal. Reward functions can be easy to construct if the problem has clearly defined goals, such as achieving a high score in a game, but becomes much more complex when goals cannot be easily defined by a scalar value.

In this study, the goal of the RL agent is to optimise the performance indicators for the *GSE*. Communicating a multi-objective (5 performance indicators) goal is significantly more complex for RL agents compared to MOEAs. This because all objectives need to be mapped to a single reward function value, contrarily this is trivial for MOEAs, as they are able to consider each objective individually when comparing solutions. A common method for multi-objective RL used in this study is the Weighted Sum method [49], which converts the multi-objective problem into a single objective problem using a scalarisation function. These functions assign a weight to each objective and the sum of each objective multiplied by its weight is used as the final value, mapping the multiple objectives into a single objective.

Three different reward functions using the Weighted Sum method were developed and evaluated for the RL agent. Each reward function was designed as a dense reward function, by providing a non-zero reward to the RL agent at each timestep, as opposed to returning the quality of the agent’s actions at the end of each simulation. Dense reward functions have been cited to better guide RL agents, and increase the speed in which an RL agent learns an optimal policy, due to the more frequently received reward (feedback) [16].

The first reward function (RF1) combined the normalised values of each of the performance indicators. At each timestep, the current values of the performance indicators were min-max normalised, using the minimum and maximum values of the performance indicators, drawn from a sample of 10^6 experiments (see Appendix E.1). As previously stated in this study, the true maximum and minimum bounds of the performance indicators cannot be known, but an approximation has been made for this reward function. Let s_t be the state of *GSE* at timestep t. The weight values in w represent an equal weighting between all performance indicators, and the sign of the weight refers to whether the performance indicator is desired to be maximised or minimised. The reward function is defined as:

$$\begin{aligned}
r(s_t, a_t) &= \sum_i w_i f_i \\
w &= \langle -1, 1, -1, -1, -1 \rangle \\
f &= \langle s_t^{ghge}, s_t^{renew}, s_t^{wholesale}, s_t^{tariff}, s_t^{unmet} \rangle
\end{aligned}$$

The second reward function (RF2) omitted the use of approximation for performance indicator ranges, instead focusing on the change in the performance indicators between timesteps. The aim was to provide feedback to the RL agent on whether it was moving in the right direction with respect to the change in performance indicators. However, the dynamics of each performance indicator vary over the simulation period. For example, towards the end of each simulation, *Unmet Demand Days* deteriorates whereas *Renewable Market Share* improves. These varying dynamics can potentially provide noise to the reward function, where one performance indicator's deterioration hides the improvement of another. The reward function is defined as:

$$\begin{aligned}
r(s_t, a_t) &= \sum_i w_i \times \frac{f_i - g_i}{g_i} \\
w &= \langle -1, 1, -1, -1, -1 \rangle \\
f &= \langle s_t^{ghge}, s_t^{renew}, s_t^{wholesale}, s_t^{tariff}, s_t^{unmet} \rangle \\
g &= \langle s_{t-1}^{ghge}, s_{t-1}^{renew}, s_{t-1}^{wholesale}, s_{t-1}^{tariff}, s_{t-1}^{unmet} \rangle
\end{aligned}$$

There are periods in the simulation period where some of the performance indicators improve or deteriorate, regardless of any action taken (e.g. *Unmet Demand Days* deteriorates after a large generator closes). This presented the risk in RF2 that the reward the agent received did not truly reflect the quality of its actions. The third reward function (RF3) sought to remove this potential skew, by assessing if the RL agent has improved the performance indicators relative to a fixed baseline generated by the BAU scenario. The values of the performance indicators at each timestep were recorded from a model simulation using the BAU scenario, and no policy actions implemented. Let b_t be the state of *GSE* at timestep t in the BAU scenario:

$$
\begin{aligned}
r(s_t, a_t) &= \sum_i w_i \times \frac{f_i - g_i}{g_i} \\
w &= \langle -1, 1, -1, -1, -1 \rangle \\
f &= \langle s_t^{ghge}, s_t^{renew}, s_t^{wholesale}, s_t^{tariff}, s_t^{unmet} \rangle \\
g &= \langle b_t^{ghge}, b_t^{renew}, b_t^{wholesale}, b_t^{tariff}, b_t^{unmet} \rangle
\end{aligned}
$$

8.4.1 Experimental Setup

We seek to evaluate the behaviour that each reward function elicits from an RL agent and to identify which reward function is able to most clearly communicate the overall goal of optimising the performance indicators. Three RL agents were trained using the DQN algorithm, with each agent using a different reward function. The agents were trained over 10^6 episodes (simulation runs), and the uncertain input parameters of the model were the values defined by the BAU scenario, resulting in the same runtime conditions for each episode. The analysis of the results will assist to determine which reward function will be used when the final RL agents are trained and evaluated using multiple scenarios (deep uncertainty) to answer the research questions of this study. The training was completed on a 44 core 2.0 GHz Intel Xeon server with multiprocessing, and took 32.61 hours (Appendix E.2). Once each agent was trained, they were tested in the BAU scenario, and their performance and policy actions were recorded. In addition, the runtime results that occur when no policy actions are implemented in any year will be included in the results as a comparative baseline.

8.4.2 Results & Discussion

The runtime results from the first policy action in 2022 for the three RL agents and the baseline data are plotted in figure 8.1, and the policy actions each agent chose are recorded in table 8.1. The results of each agent will be referred to by their reward function for this discussion. Both RF1 and RF2 appeared to have learnt “blunt” policies, using few, but high-impact actions. RF1 only used the merit-order by emission levels policy action and RF2 implemented a carbon tax from 2022 to 2029, and then used the top emitting generators 5% capacity reduction action for the rest of the simulation. The initial economic impacts of these policy actions are severe, with notable spikes in the *Wholesale Prices* and *Tariff Prices*, and a very large increase in *Unmet Demand Days* for RF2 when it implements its 5% reduction policy action. The steep increase in *Unmet Demand Days* after reducing the capacity of the top 1% emitting generators suggest the top 1% generate a significant proportion of the total electricity in the market during this time.

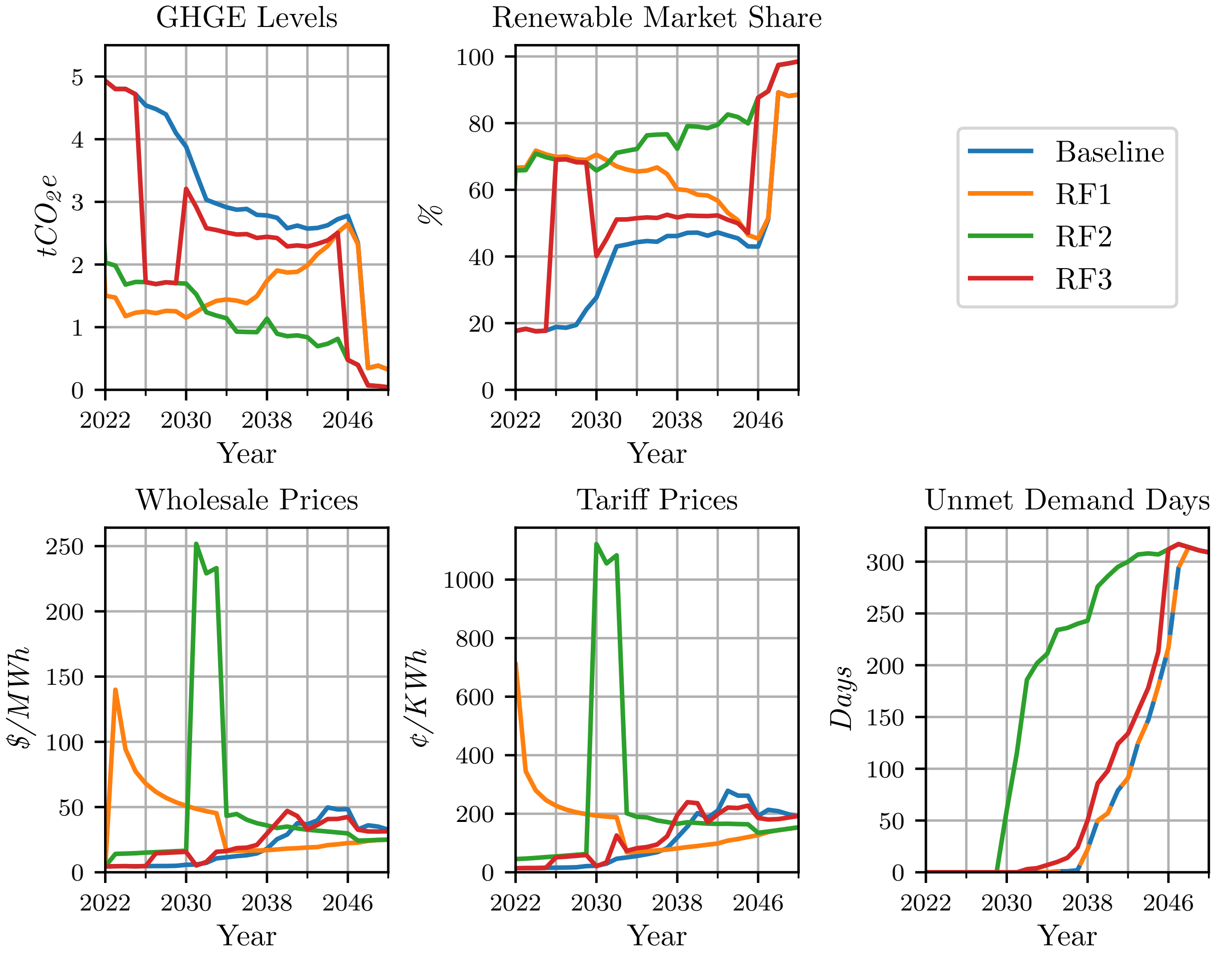


Figure 8.1: Real-time values of performance indicators during evaluation.

Conversely, the RF3 appears to have developed a less extreme, more nuanced policy for the RL agent. It begins by increasing the learning rate of solar and wind powered generators, thereby decreasing their LCOE. The following carbon tax increases the LCOE of non-renewable generators, reducing their bidding power, allowing for renewable generators to be more competitive, and have more opportunity to have successful bids, thereby reducing their LCOE. Following this, the highest emitting generator has its capacity reduced, but at this time in the *GSE*, this only has minimal impact on the economic indicators. An interesting policy action is the opening of the secondary electricity market, allowing for more renewable generators to supply electricity. Finally, it implements no policy action from 2038 to 2045, and then reduces the capacity of the top 1% emitting generators from 2046 onwards. The behaviour from 2038 onwards is a very clever strategy learned by the agent. In these years, there are only marginal gains to be made in the environmental indicators, which come at a high cost for the economic indicators. The effects from earlier policy actions that benefited renewable generators are still present in these later years, and as a result, the RF3 environmental results still outperform the baseline results.

Tabular data for the average value of the performance indicators from 2022 to 2050 can be found in table 8.2, and their percentage change from the baseline in table 8.3. All reward functions were able to improve the environmental indicators, but with varying trade-offs for the economic indicators. For example, relative to the baseline, RF2 was able to decrease the mean *GHGE Levels* by 64.50%, but increased the mean *Wholesale Prices* by 137.9%. The severity of the environmental and economic trade-offs for RF1 and RF2 were deemed too significant, suggesting these reward functions were not able to effectively encode the multi-objective nature of the problem, and hence won't be considered for the rest of this study. The RF3 reward function, whose results were less pronounced, can be considered to have effectively communicated to the RL agent the multi-objective nature of this problem, and will be used for the remainder of this study.

Year	**RF1**	**RF2**	**RF3**
2022	EmMerit	CbnTax	LR15%
2026	EmMerit	CbnTax	CbnTax
2030	EmMerit	R5%	R20%
2034	EmMerit	R5%	SecMkt
2038	EmMerit	R5%	NoAction
2042	EmMerit	R5%	NoAction
2046	EmMerit	R5%	R5%

Table 8.1: Policy actions used by each reward function during evaluation. See table 4.1 for details on each policy action.

Performance Indicator	**Baseline**	**RF1**	**RF2**	**RF3**
GHGE Levels (tCO_2e)	3.07	1.51	1.09	2.32
Renewable Market Share (%)	41.73	65.38	76.86	55.88
Wholesale Prices ($\$/MWh$)	20.28	37.42	48.27	22.93
Tariff Prices (¢/KWh)	111.24	167.90	229.84	122.52
Unmet Demand Days ($Days$)	75.86	75.86	185.14	91.86

Table 8.2: Average performance indicator values for each reward function.

Performance Indicator	**Baseline**	**RF1**	**RF2**	**RF3**
GHGE Levels (tCO_2e)	+0.00%	-50.81%	-64.50%	-24.43%
Renewable Market Share (%)	+0.00%	+57.14%	+83.33%	+33.33%
Wholesale Prices ($\$/MWh$)	+0.00%	+84.52%	+137.97%	+13.07%
Tariff Prices (¢/KWh)	+0.00%	+50.93%	+106.63%	+10.14%
Unmet Demand Days ($Days$)	+0.00%	+0.00%	+144.05%	+21.09%

Table 8.3: Percentage change from baseline results for mean performance indicator values.

8.5 Training For Deep Uncertainty

Having formalised all of the required components for the MDP, this study can begin the design and training of the RL agents that will be used to implement policy actions that optimise the performance indicators of the *GSE* under deep uncertainty. An implementation of the DQN algorithm provided the Python RL library, RLlib [100], will be used to train the RL agents. RLlib provides in-built features for parallel RL training, which leverages multiple processors on a single machine to reduce the wall time required to train an agent for a fixed number of episodes. This study will also use another Python RL library, OpenAI Gym [98], that facilitate interaction between the RL agent and the *GSE* via state/action spaces and a reward function. OpenAI Gym acts as a wrapper around computational models, and exposes APIs that allow for interaction with the underlying model as if it were an MDP. In effect, OpenAI Gym is used in this study as a medium to control the interaction between the RL agents and the *GSE*.

The first agent (A1), the Pure RL agent, will learn a policy for the *GSE* MDP using only the DQN algorithm provided by RLlib. This agent will assist to answer the first research question (section 1.2), regarding the efficacy of RL algorithms to regulate electricity markets. In addition, the results of agent A1 will provide a baseline to help assess the influence of a *Metro Map* on RL agents. No hyperparameter tuning was completed for any of the DQN agents in this study, but readers may look to the RLlib DQN documentation[1] for the default hyperparameter list. The γ discount factor is also included in the hyperparameter list. The second agent (A2), will have its actions determined by the paths through the *Metro Map*. Previously in this paper (section 3.1), we stated the use of the *Metro Map* is to effectively pre-process the actions available to agent A2. An alternative interpretation, is that the *Metro Map* pre-processes the trajectory space of the agent during its training.

The term "trajectory" is used in RL to describe the sequence of environment states and actions an agent experiences throughout an episode. Consider a simple RL agent, that always uses the same action when interacting with the *GSE*. The trajectory of that agent when the *GSE* is using the BAU scenario will be distinct from the trajectory of every other scenario, despite using the same actions. This is due to the values encoded in state space being different between scenarios, as the influence of the uncertain parameters cause these values to change. As a result, trajectories are dependent on both the scenario and policy actions used when running the *GSE*.

[1] https://docs.ray.io/en/master/rllib-algorithms.html

In the robust optimisation performed in chapter 7, each policy pathway was tested against the 500 scenarios, which resulted in 500 trajectories per policy pathway, whose performance was then evaluated by the robustness metric. The ϵ-NSGAII can therefore be thought to have policy pathways with Pareto-optimal trajectories for the 500 training scenarios, with respect to the robustness metric used. As the same 500 scenarios will be used during the training of A2 (detailed in section 8.5.2), the number of possible trajectories therefore will be equal to the number of policy pathways in the *Metro Map*, multiplied by the number of scenarios ($53 \times 500 = 26500$). This is multiple orders of magnitude less than the number of possible trajectories for agent A1 ($10^7 \times 500 = 5 \times 10^9$), and will ideally reduce the complexity of learning which policy pathways in the *Metro Map* are the best to follow given the state of the *GSE*.

To ensure that agent A2*'s* action follow a valid path in the *Metro Map*, the actions available to agent A2 need to be parameterised according to the position of the agent in the *Metro Map*. RLlib's DQN algorithm natively supports this by allowing for parametric action spaces[2] as a flag in its hyperparameters. At runtime, the OpenAI Gym program loads the structure of *Metro Map* into memory, and at the beginning of each training episode, it sets the current position of the agent in the *Metro Map* to the black root node (see figure 7.12). A bit mask of length 10 is computed (the number of possible actions), where bits are set to one if the corresponding action for that bit is valid from the root node of the *Metro Map*. In the final step of the DQN algorithm, the bit mask is used to set the probability of invalid actions being selected to zero, ensuring they are never chosen. Once the action has been chosen, the OpenAI Gym records that action, and updates the position of the agent in the *Metro Map* and the bit mask to reflect the next set of available actions. The agent is still only aware of the *GSE* via the state space, but the OpenAI Gym records the "location" of where the agent would be in the *Metro Map* as if the agent was traversing the *Metro Map*.

In addition to agent A2 having its action space parameterised, this study finds the agent should also have knowledge of the structure of the *Metro Map* to help the agent traverse the pathways. Granting the agent structural knowledge of the *Metro Map* will accommodate a greater understanding of the potential trade-offs between multiple pathways, and a more holistic knowledge of the *Metro Map*. This structural knowledge provides a symbiotic relationship between the *Metro Map* and RL agent. The *Metro Map* helps the agent by providing a set of robust policy pathways as guides, and the agent helps the *Metro Map* by providing a new way it can contribute to the policy design process, that mitigates the drawbacks of a *Metro Map*'s complexity. The complexity and the number of pathways in a *Metro Map* hinders its utility as a policy design tool for human policy-makers, but we aim to show RL agents can leverage this complexity to promote better results. To encode the structure of the *Metro Map* to the RL agent, each node in the *Metro Map* is given a unique identifier integer and passed as an additional variable to the state space, so that the agent would also know its current position in the *Metro Map*. This augmentation of the state space means agents A1 and A2 are no longer equivalent in their observation of the *GSE*, however, this small difference does not invalidate this study's intentions to compare the two agents.

[2] https://docs.ray.io/en/master/rllib-models.html#variable-length-parametric-action-spaces

8.5.1 Random Baseline

A common practice in RL studies is the use of a simple "random" baseline agent, which represents an RL agent that has undergone no training. Random agents are used to assert that a trained RL agent has learned something during its training phase that positively contributes to the overall objective, and is able to improve upon taking random actions. If a random agent performs similarly to a trained agent, then the trained agent can be considered to have not learned anything useful during training. The random agent (A1-R) will pseudo-randomly choose an action, where each action has an equal chance of being selected. The performance of agent A1-R will assist to assert if either agents, A1 or A2, were able to learn useful strategies during their training phase, and will confirm if RL is able to regulate the *GSE* to optimise the performance indicators.

Given the pathways in the *Metro Map* are robustly optimised for the 500 training scenarios, there is a valid argument that there is no need to try to "learn" how to traverse the map. Any pathway in the *Metro Map* could be chosen, and would still yield relatively strong results. To validate the use of combining the *Metro Map* with RL algorithms, we consider a second random baseline agent (A2-R). Agent A2-R pseudo-randomly chooses one of the Pareto-optimal policy pathways at the start of each episode. This random baseline will assist in demonstrating if the RL agent is able to leverage the benefits of the *Metro Map*, particularly how it combines MOEA robustly optimised pathways, and allows the agent to still be adaptive by branching nodes in the *Metro Map*.

8.5.2 Multi-Phase Training

Training the two RL agents to optimise the performance indicators in the presence of deep uncertainty was a difficult task. Deep uncertainty was to be represented in the same way it was in the MOEA robust optimisation (chapter 7), by using the set of 500 different scenarios generated for the ϵ-NSGAII to influence the runtime conditions of the *GSE*. At the start of each training episode, the agent would pseudo-randomly select one of the 500 scenarios to input to the *GSE*. This exposed the agent to a wide variety of scenarios during training, which would hopefully enable the DQN algorithm to develop a strategy (policy) that could adapt to deep uncertainty. Unfortunately, this training method produced severely undesirable results, with neither RL agents producing better results than the random baselines. We suspected this level of uncertainty overwhelmed the DQN algorithm, stopping it from ever being able to learn by interaction with the model. Even for 10^6 training episodes (29×10^6 years of simulated experience), neither agent exhibited behaviour (e.g. mean episode reward) that represented a successful RL training phase.

Inspiration was taken from an RL concept known as curriculum learning [149] to design a more balanced and progressive method to expose deep uncertainty whilst training the RL agents. Curriculum learning is a method for designing the training phase of an RL agent, that optimises the order in which an agent gains new experience. Studies on curriculum learning have shown commencing training with simpler problems, and gradually increasing the difficulty can increase the speed of training, and the performance of the agent [149]. Employing a curriculum in algorithmic training has been done in many problem domains, such as robotics, grammar learning, and classification [149]. A very simple example is training an agent to win in the game of chess. First, the agent is trained to win on a board with only pawns. Next, a new piece (e.g. rook) is added to the game, the agent continues training. New pieces are periodically added throughout training until all chess pieces are present on the board. In effect, we seek to start off with a simple problem, and progressively increase the problem difficulty.

As it was suspected the extent of deep uncertainty introduced by the 500 different scenarios was the driving force for the failure of the agents to learn, we sought to change the way the agent is exposed to uncertainty. By progressively introducing uncertain scenarios, we aim to progressively make the task of the RL agents harder. The training phase was subsequently divided into two parts. Initially the agents only had access to 250 scenarios to pseudo-randomly choose from at the beginning of each training episode. The agents initially trained with these 250 available scenarios for 10^6 episodes, and then were able to select from all 500 scenarios for an additional 10^6 episodes. The mean episode reward of both agents during training is plotted in Appendix E.3. Designing the training in two phases provided a progressive introduction of uncertainty in the *GSE*, which did not overwhelm the DQN algorithm such that it was not able to operate as intended. The training was completed on a 44 core 2.0 GHz Intel Xeon server with multiprocessing, and took 94.16 hours.

8.6 Evaluation

The RL agents' performance will be based upon the performance indicator results from a set of $10,000$ scenarios sampled using Latin Hypercube Sampling [128]. A no-action baseline similar to the one in the ϵ-NSGAII analysis (chapter 7) is also included for this analysis. The data for the baseline was generated by running the $10,000$ scenarios with no policy actions being taken at any point in the simulation. The baseline analysis is essential to validate the RL agents were able to meaningfully assist the *GSE* with the policy actions available. The results from two random RL agents are also included to better critique the value of using RL techniques for this study. Similar to the evaluation of the ϵ-NSGAII results, this analysis is guided by the recommendations of Kwakkel *et al.* [102], who cited when evaluating the performance of policy actions in deeply uncertain systems, the focus of the evaluation should be on both the values and distribution of the performance indicators. Acronyms that will be used to refer to the four RL agents in this are displayed in table 8.4. Tables containing summary statistics of the results (e.g. mean, standard deviation) evaluated in this section can be found in Appendix E.4.

Name	**Description**	**Acronym**
Pure RL	RL agent trained with the DQN algorithm.	A1
Random Pure RL	RL agent that chooses actions randomly.	A1-R
DAPP/RL	RL agent trained with the DQN algorithm that follows the *Metro Map*.	A2
Random DAPP/RL	RL agent that chooses random policy pathways in the *Metro Map*.	A2-R

Table 8.4: Description and acronyms of the RL agents used for evaluation.

8.6.1 Performance Indicator Results

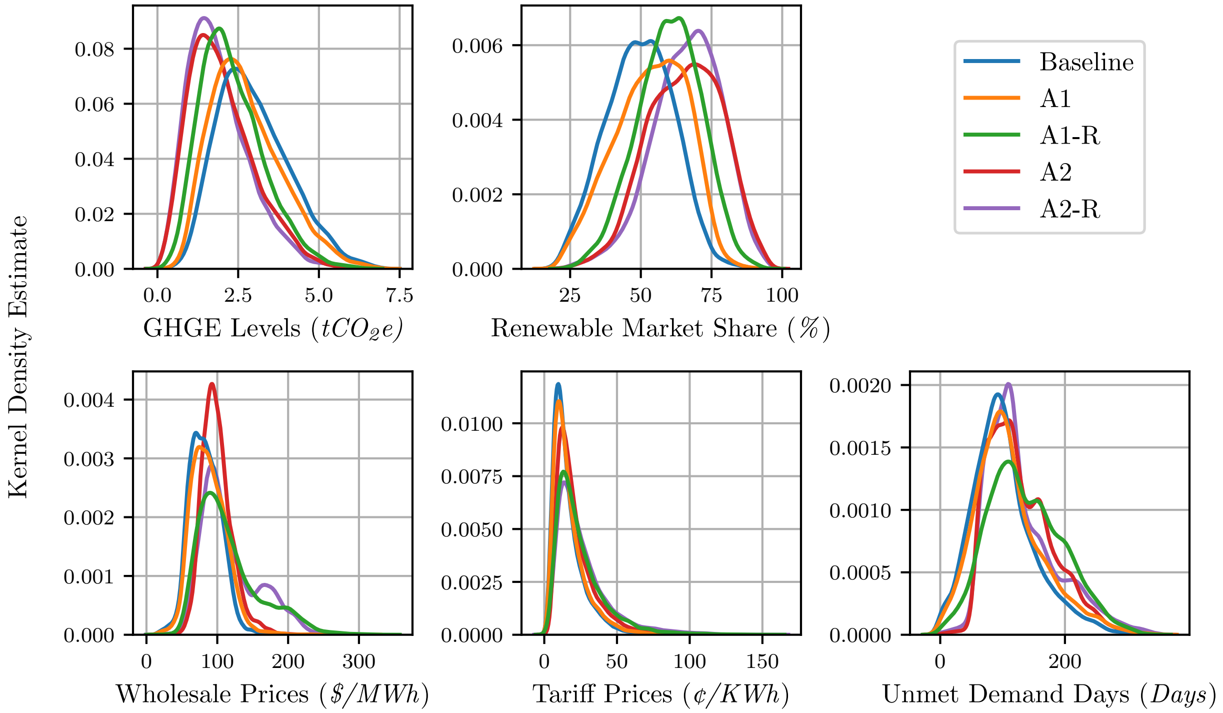


Figure 8.2: KDE plots of the results for all agents and performance indicators.

Kernel density estimates of the performance indicator results were fitted in figure 8.2. Unlike the results from the ϵ-NSGAII experimentation (chapter 7), there is no clear distinction between the results generated from the computational techniques we are evaluating (RL) and the no-action baseline results. This is unfortunate, as we hoped to see the RL agents have significantly more desirable distributions for at least the environmental indicators. Nonetheless, this isn't indicative of the agents not being as effective. They were all able to positively influence the environmental indicators, and most pleasingly, their distributions of the economic indicators are similar to the no-action baseline results. As the policy actions for this study predominantly have negative economic impacts, this preliminary insight indicates that the RL agents may have made a more subtle balance between the environmental and economic indicators.

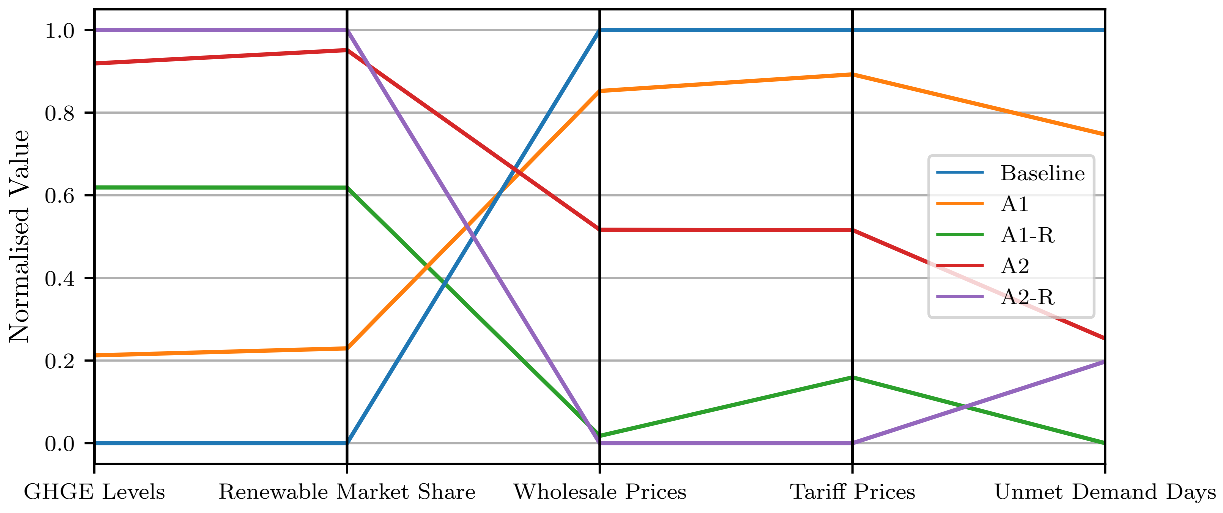


Figure 8.3: Min-max normalised mean performance indicators values for all agents.

The mean value of the results per performance indicator were min-max normalised, such that larger normalised values were desired, and were plotted in figure 8.3. As expected, the no-action baseline results strongly favoured the economic indicators due to the natural cost-minimising function of the merit-order market. Figure 8.3 starts to convey the relative priorities of the two types of RL agents, Pure RL (A1/A1-R), and DAPP/RL (A2/A2-R). Regarding Pure RL, agent A1 makes only relatively minor improvements on the environmental indicators, without causing too much detriment to the economic indicators. Conversely, agent A2 is more environmentally focused, performing strongly for the environmental indicators, and relatively average on the economic indicators. Both agents A1 and A2, use the same reward function, which can be considered as they share the same goal. The difference in their apparent prioritisation between the environmental and economic indicators is a direct result of the employment of the *Metro Map* in agent A2.

Although agent A1-R performed as economically poorly as agent A2-R, its policies did not have the environmental benefits that were displayed by agent A2-R. The only other agent that could approach agent A2-R's environmental performance was the other DAPP/RL based agent, agent A2. We contend that identifying sequences of actions (or policy pathways) that greatly benefit the environmental indicators is a difficult challenge for the *GSE*. The ϵ-NSGAII's ability to specify and optimise multi-objective problems by analysing each dimension of the objective space allowed it to identify policy pathways of great environmental benefit, and such knowledge was then encoded into the DAPP/RL agents. Conversely, using only RL techniques, or taking random actions, was not enough to yield strong environmental results for the *GSE*.

Comparing the results between agents A1 and A1-R, using a random policy has lead to relatively poor results for all performance indicators, and a differently skewed trade-off between the environmental and economic indicators. This suggests that the use of the DQN algorithm to optimise the performance indicators has merit, and that it was indeed able to make tangible differences, compared to taking random actions.

A final insight from figure 8.3 comes from the economic differences of the DAPP/RL agents. Both agents A2 and A2-R possess similar environmental results relative to all other results, but economically, agent A2-R diverges to the bottom of all results. This will be discussed in greater detail in section 8.7, but this is indicative of how RL techniques have been applied to better leverage the robustly optimised policy pathways generated by the ϵ-NSGAII.

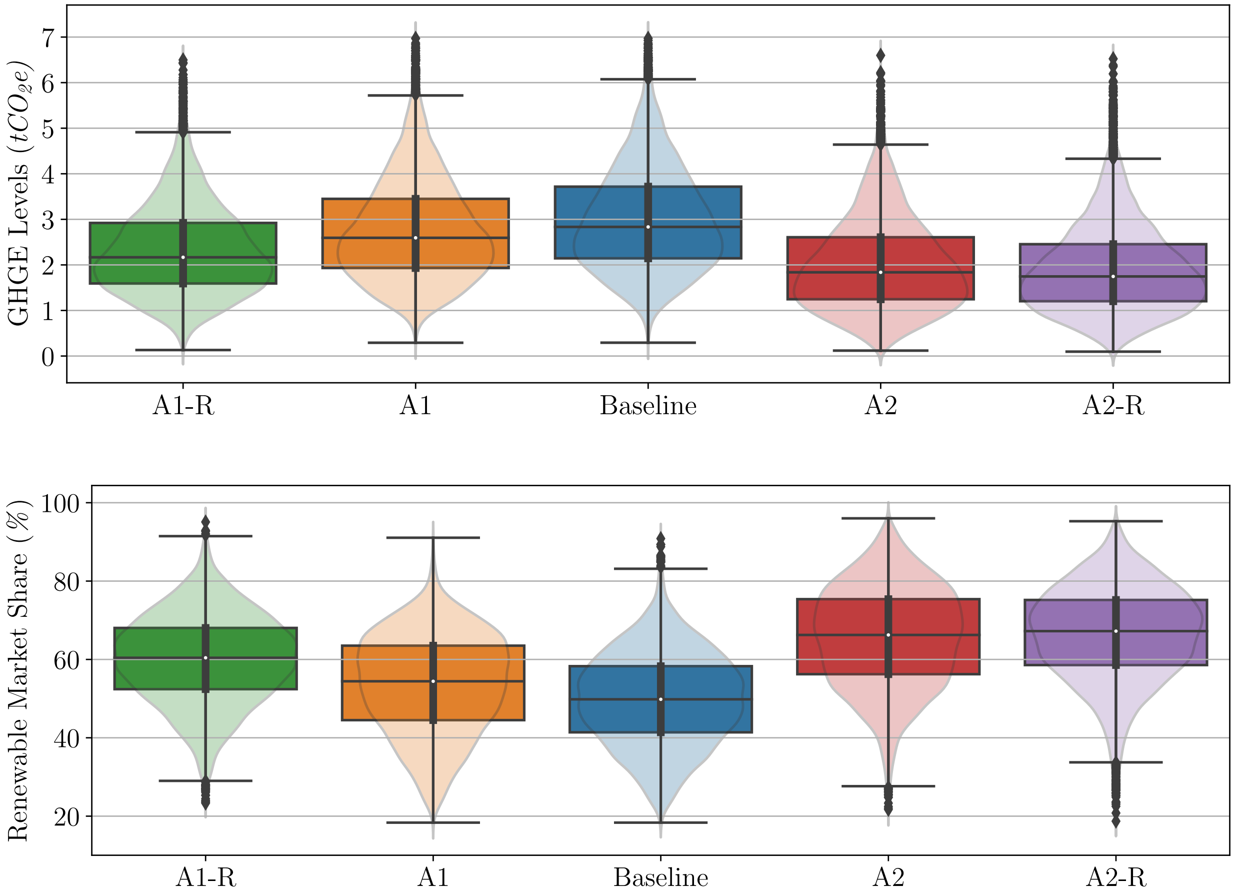


Figure 8.4: RL agent results for 10,000 scenarios - Economic Indicators.

As demonstrated from figure 8.4, agent A2 was able to make considerable improvements to the environmental indicators, reducing the mean *GHGE Levels* from the baseline value of $2.99tCO_2e$ to $2.00tCO_2e$, and increasing the *Renewable Market Share* from 49.74% to 65.65%. Whereas, the agent A1's influence was comparatively smaller, reducing the mean *GHGE Levels* by $0.23tCO_2e$ and increasing *Renewable Market Share* by 3.82 percentage points. Given the potential power of the policy actions in this study to benefit the environmental indicators, as shown by agents A2, A2-R, and to lesser degree agent A1-R, it is not promising to see such minute environmental impacts by agent A1.

Agents A2 and A2-R have similar environmental distributions, despite utilising the pathways through the *Metro Map* in a very different manner. In chapter 7, we discussed the environmental indicators values of the robustly optimised policy pathways approached a potential upper bound, explaining the similarity in environmental results for all metrics. We find this upper -bound also explains the similar environmental results of agents A2 and A2-R.

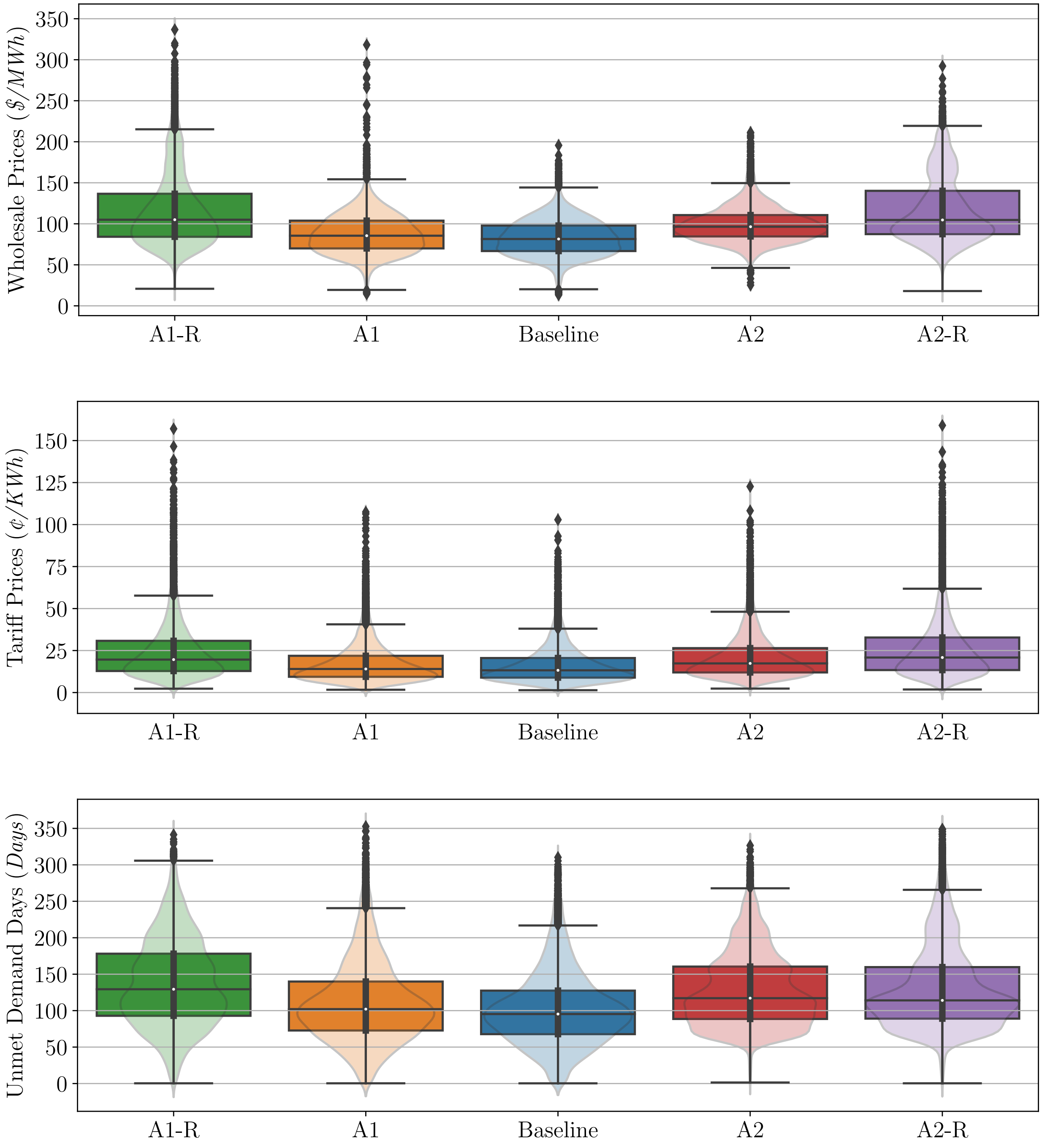


Figure 8.5: RL agent results for $10,000$ scenarios - Economic Indicators.

Concerning the economic indicators, most impressively, agent A2 was able to generate the smallest *Wholesale Prices* standard deviation of 21.38*$/MWh*, smaller than the baseline, and did not have any extreme outliers, unlike the other agents. The two random agents both demonstrate their inability to optimise the economic indicators. Whilst poorer performance of a purely random RL agent such as agent A1-R, it was expected the utilisation of the *Metro Map* by agent A2-R would result in it still having relatively sound results for all indicators. As the policy pathways A2-R uses were also optimised for all performance indicators, we expected to see strong economic performance, such as the results of agent A2. We determine that agent A2-R's poor economic performance compared to agent A2 is due to the use of the DQN algorithm. The adaptive capability of agent A2 allowed it to observe the *GSE* during runtime, and choose desirable policy actions at branching points in the *Metro Map* that were best suited to its observation of the *GSE*.

Agent A2 has shown it is able to use the Pareto-optimal paths that make up the *Metro Map*, in such a manner that it can yield more desirable results, regardless of the deep uncertainty it faces. We also contend the reward function used, which prioritised the economic indicators for agent A1, may also explain agent A2's strong economic performance. It is evident that randomly choosing a path through the *Metro Map* leads to strong environmental results, but when navigating through the map, with a reward function that is economically oriented, we produced desirable results for all environmental and economic indicators.

The economic results of agent A1 are unfortunately very similar to its environmental results. This is likely a reflection on the reward function agent A1 uses. Without the guidance of the *Metro Map*, when exposed to deep uncertainty, the agent was not able to effectively find solutions to the multi-objective problem the reward function attempted to communicate. However, its economic performance was much stronger than its random equivalent, agent A1-R, affirming the notion of the economic priority of the reward function used in this study, and the capability of the DQN algorithm to learn strategies to optimise some of the performance indicators for this study.

8.7 Discussion

Despite the extensive research and experimentation that went into designing and training agent A1, it was not able to notably influence the market dynamics of *GSE* to optimise the performance indicators. Relative to the baseline results, it made small environmental improvements at a small economic cost. The actions agent A1 used could be considered to have been chosen according to a very risk-averse strategy, hence the minimal changes relative to the baseline results. In addition, when compared to its random equivalent, agent A1-R, it clearly demonstrated it had learned a policy that supported the overall aim of the agent, and was superior to random action choices. It is curious though that agent A1's results were so similar to the baseline, given agent A1 used $2,712$ different sequences of actions (policy pathways) for the $10,000$ evaluations scenarios. This further highlights the challenges of finding robust, well performing policy pathways in the *GSE*.

Overall, agent A1 took a subdued approach to optimising the performance indicators. When considering **RQ1** of this study, whether RL techniques can assist to regulate an electricity market to promote transitioning to sustainable electricity sources, we conclude that agent A1 was not able to effectively perform this task. Due to the degree of deep uncertainty in the *GSE*, and the five different performance indicators used to guide the agent, the task of regulating the market is too challenging to be performed by RL techniques alone.

In contrast to agent A1, the leveraging of the *Metro Map* by agent A2 led to strong results for all performance indicators. Even when faced with a system under deep uncertainty, agent A2 was able to use the *Metro Map* to utilise the robustly optimised policy pathways to generate sounds results. The values of the environmental indicators from the evaluation scenarios demonstrated a clear, tangible improvement across all scenarios. As seen many times in this study, environmental performance comes at a trade-off of economic performance, but agent A2 was able to generate tight economic distributions, which is a key strength when designing policies in deep uncertainty literature.

A potential concern in the behaviour of agent A2 is that only 3 policy pathways in the *Metro Map* were used for 30% of the $10,000$ evaluation scenarios. We suspect a key reason the agent used those few pathways for such a large share of the evaluations structures is due to the structure of the *Metro Map*. Many nodes in the *Metro Map* study have no branching factor, such that there is no choice on the next policy action to activate, as there is only one path out of that node. In effect, the agent was locked into a given policy pathway early in the *GSE* simulation. To address this issue, future work could ensure that more policy pathways are identified using an MOEA to construct the *Metro Map*, or less policy actions are available. Either of these recommendations would ensure that there is a higher probability that policy pathways identified by the MOEA share a "prefix", and as a result could be combined in the *Metro Map*.

The analysis in this chapter demonstrates how agent A2 has been able to assist in regulating the *GSE*, promoting the transition to sustainable electricity sources. The true task of agent A2 is to answer **RQ2**, how the *Metro Map*, an Exploratory Modelling technique generated by MOEAs, can assist RL techniques when faced with deep uncertainty. The superior performance of agent A2 relative to all other agents evaluated, clearly conveys the strength and validity of combining the *Metro Map* with RL techniques, thus providing sound evidence to address **RQ2**. The structure of the *Metro Map*, and the adaptive capabilities of RL techniques resulted in an agent who produced strong results in the face of deep uncertainty.

The reason for developing agent A2-R was to explore an alternate consideration of the proposed DAPP/RL technique for this study. Comparing agents A1 and A2 showed how the combined DAPP/RL could outperform Pure RL, but there was still a question of if there was any merit in training an RL agent to use the *Metro Map*, rather than just using any of the Pareto-optimal robustly optimised policy pathways identified by the ϵ-NSGAII. Initially this study was concerned that given these policy pathways were already robustly optimised, and Pareto-optimal with respect to the *90-P* robustness metric, agent A2-R would perform strongly across all scenarios, and would not benefit from the use of RL techniques. However, comparing the results between agents A2 and A2-R (particularly for the economic indicators), despite the policy pathways of both agents being robustly optimised, some policy pathways are better suited to some scenarios than others, and hence their ability to optimise the performance indicators greatly differed. By utilising RL techniques, agent A2 has been able to use the policy pathways in the *Metro Map* that are best suited to the scenario of the *GSE* it faced.

Chapter 9

Conclusion

9.1 Future Work

Despite the sound results of the proposed DAPP/RL based agent, three key aspects of this work that would benefit from further studies have been identified. Firstly, the evaluation of the RL agents conducted in section 8.6 suggested that only using RL techniques for the policy design process was ineffective, resulting in insignificant changes of the performance indicators relative to our no-action baseline. During the design phase of the RL reward function, the agent was shown to make significant improvements to the performance indicators, but only for when the *GSE* was using the BAU scenario. Accordingly, we reason that in the absence of deep uncertainty, RL is an effective technique for policy design. The deep uncertainty of the *GSE* in the final evaluation of the agent overwhelmed the agent, resulting in its poor performance. To extend this research, future studies should seek for new methods for designing the reward function, such that the reward function consistently provides meaningful feedback to the agent when faced with deep uncertainty, and further, integrate additional techniques from existing multi-objective RL literature.

The second area for further investigation pertains to the time span for which the DAPP/RL agent is available in policy design. A policy pathway, or even a *Metro Map* are inherently only useful in practice if they are followed by the policy-maker. If an “unavailable” action is chosen whilst following the *Metro Map*, the remainder of the *Metro Map* can be discarded, as there is no path that reflects the policy-makers actions. The proposed DAPP/RL technique of this study possesses the same weakness as *Metro Maps*. If a policy-maker deviates from the recommendations of the agent to use an action that is not valid within the *Metro Map*, there is no longer a position within the *Metro Map* to use for the state representation when querying the agent. This means that the agent cannot be meaningfully queried for recommendations on what actions to take.

If the location of an RL agent within the *Metro Map* was not required, it can continually be employed as an additional consultant for policy-makers, to advise what action to take, and by the Markov property, it can still be effective no matter what past actions were taken, or how the future has unfolded. As a result, this study highly recommends further investigation into different ways the RL agent can learn from, and leverage the robustness and multi-objective capabilities of the policy pathways within the *Metro Map*, with particular emphasis to remove the dependency of the agent's state space on the *Metro Map*. Future studies may find the literature curriculum learning [149] and transfer learning [150] as relevant fields in the literature to begin exploring this idea.

Our final recommendation is to extend the evaluation methodology of section 8.6, by comparing the quality of the DAPP/RL agents against human policy-makers. In this study we used random action and no action baseline agents for assessing our agents' performance in the *GSE*. Whilst these baselines helped assert the efficacy and validity of the designed RL agents, in a traditionally heavily human involved process such as policy design, our RL agents quality compared to human policy-makers is an overarching question that should be answered. To better understand how RL agents fare against human-policy makers, investigation should be conducted to compare the results of the RL agents, against human policy-makers following traditional policy design paradigms, such as DAP [24] or MORDM [47]. This additional comparison would help consolidate the efficacy and validity of using RL techniques for policy design problems, for systems under deep uncertainty.

9.2 Concluding Remarks

This work has explored novel policy design methods for systems under deep uncertainty, using the Victorian electricity network, within the Australian National Electricity Market as a case study. Using a computational model of the market known as the *GSE* [92], we explored how Reinforcement Learning can assist in designing adaptive policies to promote a transition to a more sustainable future, with lower greenhouse gas emissions, greater renewable electricity use, whilst minimising economic impacts. Guided by the research questions posed in section 1.2, we investigated if Reinforcement Learning was able to regulate the market to promote this transition, and whether an existing policy design paradigm, Dynamic Adaptive Policy Pathways, could be combined with Reinforcement Learning to improve the quality of the policies designed by the RL agent.

To the best of our knowledge, our work on using RL techniques to act as an electricity market regulator, rather than a market participant, is the first of its kind, addressing an existing gap in the literature. In addition, we have combined the strengths of both Exploratory Modelling and RL techniques, by utilising the *Metro Map* developed by the Dynamic Adaptive Policy Pathways as a guide for an RL agent. Our final results demonstrated our DAPP/RL technique was able to greatly improve the quality of adaptive market policies when compared to another RL agent that only employed RL techniques. In particular, the adaptive structure of the *Metro Map* allowed an RL agent to better grapple the challenges of policy design under deep uncertainty, in a multi-objective problem domain.

The greatest significance of this work lies in the contribution of using RL agents for policy design problems for systems under deep uncertainty. Traditional optimisation tools for policy design, such as Multi-Objective Evolutionary Algorithms, output a set of Pareto-optimal solutions (policy pathways), where the knowledge the MOEA gained from the *GSE* was reflected in the policy pathways, but still required human policy-makers to further evaluate the output pathways to be able to leverage any of the findings from the MOEA. By using RL techniques, we implemented an intelligent policy design agent, which could be considered as an artificial policy-maker.

By making the agent repeatedly interact with the *GSE* over millions of timesteps, the aim is for the agent to learn the dynamics of the model such that it becomes an “oracle” on the model, knowing what policy actions to take in any scenario to optimise the performance indicators. Unlike past studies that have created computational techniques for policy design under deep uncertainty [151], we have contributed RL agents to be consultants for policy design, rather than tools. The distinction between a tool and consultant is that a tool assists policy-makers to come up with a solution, whereas the consultant returns a definitive answer on what policy action to use when queried. An RL agent can be considered an expert consultant, an artificial policy-maker who has gained intricate knowledge on the problem through experience, and can advise the best policy action to take. We seek for this work to be a seminal piece of literature to promote further investigation into the use of Reinforcement Learning as a computational tool to assist in the policy design process.

Bibliography

[1] C. Zou, Q. Zhao, G. Zhang, and B. Xiong, “Energy revolution: From a fossil energy era to a new energy era,” *Natural Gas Industry B*, vol. 3, no. 1, pp. 1–11, Jan. 2016, ISSN: 23528559. DOI: `10.1016/j.ngib.2016.02.001`.

[2] J. Köhler *et al.*, “An agenda for sustainability transitions research: State of the art and future directions,” *Environmental Innovation and Societal Transitions*, vol. 31, pp. 1–32, Jun. 2019, ISSN: 22104224. DOI: `10.1016/j.eist.2019.01.004`.

[3] E. A. Moallemi, F. de Haan, J. Kwakkel, and L. Aye, “Narrative-informed exploratory analysis of energy transition pathways: A case study of India’s electricity sector,” *Energy Policy*, vol. 110, pp. 271–287, Nov. 2017, ISSN: 03014215. DOI: `10.1016/j.enpol.2017.08.019`.

[4] E. Trutnevyte *et al.*, “Societal transformations in models for energy and climate policy: the ambitious next step,” *One Earth*, vol. 1, no. 4, pp. 423–433, 2019, ISSN: 2590-3322.

[5] B. KROPOSKI, “Integrating high levels of variable renewable energy into electric power systems,” *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 6, pp. 831–837, 2017, ISSN: 2196-5420. DOI: `10.1007/s40565-017-0339-3`. [Online]. Available: `https://doi.org/10.1007/s40565-017-0339-3`.

[6] S. Pfenninger, A. Hawkes, and J. Keirstead, *Energy systems modeling for twenty-first century energy challenges*, May 2014. DOI: `10.1016/j.rser.2014.02.003`.

[7] H. W. J. Rittel and M. M. Webber, “Dilemmas in a general theory of planning,” *Policy Sciences*, vol. 4, no. 2, pp. 155–169, 1973, ISSN: 1573-0891. DOI: `10.1007/BF01405730`. [Online]. Available: `https://doi.org/10.1007/BF01405730`.

[8] E. A. Moallemi and S. Malekpour, “A participatory exploratory modelling approach for long-term planning in energy transitions,” *Energy Research and Social Science*, vol. 35, pp. 205–216, Jan. 2018, ISSN: 22146296. DOI: `10.1016/j.erss.2017.10.022`.

[9] R. J. Lempert, S. W. Popper, and S. C. Bankes, *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis.* RAND Corporation, 2003, ISBN: 0-8330-3275-5. DOI: `10.7249/MR1626`. [Online]. Available: `https://www.rand.org/pubs/monograph_reports/MR1626.html`.

[10] A. Q. Gilbert and B. K. Sovacool, “Looking the wrong way: Bias, renewable electricity, and energy modelling in the United States,” *Energy*, vol. 94, pp. 533–541, Jan. 2016, ISSN: 03605442. DOI: `10.1016/j.energy.2015.10.135`.

[11] J. H. Kwakkel, G. Yücel, J. H. Kwakkel, and G. Yücel, “An Exploratory Analysis of the Dutch Electricity System in Transition,” *J Knowl Econ*, vol. 5, pp. 670–685, 2014. DOI: `10.1007/s13132-012-0128-1`.

[12] E. A. Moallemi and J. Köhler, “Coping with uncertainties of sustainability transitions using exploratory modelling: The case of the MATISSE model and the UK’s mobility sector,” *Environmental Innovation and Societal Transitions*, vol. 33, pp. 61–83, Nov. 2019, ISSN: 22104224. DOI: `10.1016/j.eist.2019.03.005`.

[13] D. Cao *et al.*, "Reinforcement Learning and Its Applications in Modern Power and Energy Systems: A Review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029–1042, Nov. 2020, ISSN: 21965420. DOI: 10.35833/MPCE.2020.000552.

[14] D. Cao, J. Li, D. Cai, Q. Huang, Y. Teng, and W. Hu, "Design and Application of Big Data Platform Architecture for Typical Scenarios of Power System," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*, 2018, pp. 1–5. DOI: 10.1109/PESGM.2018.8586266.

[15] A. T. Perera and P. Kamalaruban, *Applications of reinforcement learning in energy systems*, Mar. 2021. DOI: 10.1016/j.rser.2020.110618.

[16] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[17] V. A. W. J. Marchau, W. E. Walker, P. J. T. M. Bloemen, and S. W. Popper, *Decision making under deep uncertainty: from theory to practice*. Springer Nature, 2019.

[18] J. H. Kwakkel, W. E. Walker, and V. A. W. J. Marchau, "Planning and Design," vol. 39, pp. 533–550, 2012. DOI: 10.1068/b37151.

[19] W. E. Walker, R. J. Lempert, and J. H. Kwakkel, "Deep Uncertainty," in *Encyclopedia of Operations Research and Management Science*, Boston, MA: Springer US, 2013. DOI: 10.1007/978-1-4419-1153-7{_}1140.

[20] H. R. Maier, J. H. Guillaume, H. van Delden, G. A. Riddell, M. Haasnoot, and J. H. Kwakkel, "An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together?" *Environmental Modelling and Software*, vol. 81, pp. 154–164, Jul. 2016, ISSN: 13648152. DOI: 10.1016/j.envsoft.2016.03.014.

[21] D. McInerney, R. Lempert, and K. Keller, "What are robust strategies in the face of uncertain climate threshold responses?" *Climatic change*, vol. 112, no. 3, pp. 547–568, 2012, ISSN: 1573-1480.

[22] D. B. Agusdinata, "Exploratory modeling and analysis: a promising method to deal with deep uncertainty," Jan. 2008.

[23] W. E. Walker, M. Haasnoot, and J. H. Kwakkel, "Adapt or Perish: A Review of Planning Approaches for Adaptation under Deep Uncertainty," *Sustainability*, vol. 5, no. 3, pp. 955–979, 2013, ISSN: 2071-1050. DOI: 10.3390/su5030955. [Online]. Available: https://www.mdpi.com/2071-1050/5/3/955.

[24] W. E. Walker, S. A. Rahman, and J. Cave, "Adaptive policies, policy analysis, and policy-making," *European journal of operational Research*, vol. 128, no. 2, pp. 282–289, 2001, ISSN: 0377-2217.

[25] D. Swanson and S. Bhadwal, *Creating adaptive policies: A guide for policymaking in an uncertain world*. IDRC, 2009, ISBN: 8132101472.

[26] K. Mason and S. Grijalva, "A review of reinforcement learning for autonomous building energy management," *Computers and Electrical Engineering*, vol. 78, pp. 300–312, Sep. 2019, ISSN: 00457906. DOI: 10.1016/j.compeleceng.2019.07.019.

[27] S. Bankes, "Exploratory Modeling for Policy Analysis," *Operations Research*, vol. 41, no. 3, pp. 435–449, Jun. 1993, ISSN: 0030-364X. DOI: 10.1287/opre.41.3.435. [Online]. Available: https://doi.org/10.1287/opre.41.3.435.

[28] R. A. Halim, J. H. Kwakkel, and L. A. Tavasszy, "A scenario discovery study of the impact of uncertainties in the global container transport system on European ports," *Futures*, vol. 81, pp. 148–160, Aug. 2016, ISSN: 00163287. DOI: 10.1016/j.futures.2015.09.004.

[29] C. Helgeson, "Structuring Decisions Under Deep Uncertainty," *Topoi*, vol. 39, no. 2, pp. 257–269, 2020, ISSN: 1572-8749. DOI: 10.1007/s11245-018-9584-y. [Online]. Available: https://doi.org/10.1007/s11245-018-9584-y.

[30] A. A. Watson and J. R. Kasprzyk, "Incorporating deeply uncertain factors into the many objective search process," *Environmental Modelling and Software*, vol. 89, pp. 159–171, Mar. 2017, ISSN: 13648152. DOI: 10.1016/j.envsoft.2016.12.001.

[31] J. H. Kwakkel and M. Haasnoot, "Supporting DMDU: A taxonomy of approaches and tools," in *Decision Making under Deep Uncertainty*, Springer, Cham, 2019, pp. 355–374.

[32] V. Marchau, W. E. Walker, and G. P. van Wee, "Dynamic adaptive transport policies for handling deep uncertainty," *Technological Forecasting and Social Change*, vol. 77, no. 6, pp. 940–950, 2010, ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2010.04.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162510000739.

[33] V. Marchau, W. Walker, and R. van Duin, "An adaptive approach to implementing innovative urban transport solutions," *Transport Policy*, vol. 15, no. 6, pp. 405–412, Nov. 2008, ISSN: 0967070X. DOI: 10.1016/j.tranpol.2008.12.002.

[34] R. J. Hansman, C. Magee, R. De Neufville, R. Robins, and D. Roos, "Research agenda for an integrated approach to infrastructure planning, design and management," *International Journal of Critical Infrastructures*, vol. 2, no. 2-3, pp. 146–159, 2006, ISSN: 1475-3219.

[35] W. E. Walker, V. A. W. J. Marchau, and D. Swanson, "Addressing deep uncertainty using adaptive policies: Introduction to section 2," *Technological Forecasting and Social Change*, vol. 77, no. 6, pp. 917–923, 2010, ISSN: 0040-1625.

[36] J. H. Kwakkel, W. E. Walker, and V. A. W. J. Marchau, "Assessing the Efficacy of Dynamic Adaptive Planning of Infrastructure: Results from Computational Experiments," *Environment and Planning B: Planning and Design*, vol. 39, no. 3, pp. 533–550, Jan. 2012, ISSN: 0265-8135. DOI: 10.1068/b37151. [Online]. Available: https://doi.org/10.1068/b37151.

[37] R. J. Lempert, D. G. Groves, S. W. Popper, and S. C. Bankes, "A General, Analytic Method for Generating Robust Strategies and Narrative Scenarios," *Management Science*, vol. 52, no. 4, pp. 514–528, Apr. 2006, ISSN: 0025-1909. DOI: 10.1287/mnsc.1050.0472. [Online]. Available: https://doi.org/10.1287/mnsc.1050.0472.

[38] J. A. Dewar, *Assumption-based planning: a tool for reducing avoidable surprises*. Cambridge University Press, 2002, ISBN: 0521001269.

[39] J. Rosenhead, M. Elton, and S. K. Gupta, "Robustness and optimality as criteria for strategic decisions," *Journal of the Operational Research Society*, vol. 23, no. 4, pp. 413–431, 1972, ISSN: 0160-5682.

[40] D. G. Groves, *New methods for identifying robust long-term water resources management strategies for California*. The Pardee RAND Graduate School, 2005, ISBN: 0542410249.

[41] D. G. Groves and R. J. Lempert, "A new analytic method for finding policy-relevant scenarios," *Global Environmental Change*, vol. 17, no. 1, pp. 73–85, Feb. 2007, ISSN: 09593780. DOI: 10.1016/j.gloenvcha.2006.11.006.

[42] R. J. Lempert and D. G. Groves, "Identifying and evaluating robust adaptive policy responses to climate change for water management agencies in the American west," *Technological Forecasting and Social Change*, vol. 77, no. 6, pp. 960–974, Jul. 2010, ISSN: 00401625. DOI: 10.1016/j.techfore.2010.04.007.

[43] J. R. Fischbach, *Managing New Orleans flood risk in an uncertain future using non-structural risk mitigation*. The Pardee RAND Graduate School, 2010, ISBN: 1124008799.

[44] S. Popper, J. P. Griffin, C. Berrebi, T. Light, and E. Y. Min, "Natural gas and Israel's energy future: A strategic analysis under conditions of deep uncertainty," *RAND policy report*, 2009.

[45] C. McPhail, H. R. Maier, J. H. Kwakkel, M. Giuliani, A. Castelletti, and S. Westra, "Robustness Metrics: How Are They Calculated, When Should They Be Used and Why Do They Give Different Results?" *Earth's Future*, vol. 6, no. 2, pp. 169–191, Feb. 2018, ISSN: 2328-4277. DOI: https://doi.org/10.1002/2017EF000649. [Online]. Available: https://doi.org/10.1002/2017EF000649.

[46] J. H. Hyun, J. Y. Kim, C. Y. Park, and D. K. Lee, "Modeling decision-maker preferences for long-term climate adaptation planning using a pathways approach," *Science of The Total Environment*, vol. 772, p. 145 335, Jun. 2021, ISSN: 00489697. DOI: 10.1016/j.scitotenv.2021.145335.

[47] J. R. Kasprzyk, S. Nataraj, P. M. Reed, and R. J. Lempert, "Many objective robust decision making for complex environmental systems undergoing change," *Environmental Modelling and Software*, vol. 42, pp. 55–71, Apr. 2013, ISSN: 13648152. DOI: 10.1016/j.envsoft.2012.12.007.

[48] F. Neumann and A. Q. Nguyen, "On the Impact of Utility Functions in Interactive Evolutionary Multi-objective Optimization," in *Simulated Evolution and Learning*, G. Dick *et al.*, Eds., Cham: Springer International Publishing, 2014, pp. 419–430, ISBN: 978-3-319-13563-2.

[49] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: new insights," *Structural and Multidisciplinary Optimization*, vol. 41, no. 6, pp. 853–862, 2010, ISSN: 1615-1488. DOI: 10.1007/s00158-009-0460-7. [Online]. Available: https://doi.org/10.1007/s00158-009-0460-7.

[50] J. S. Arora, "Chapter 18 - Multi-objective Optimum Design Concepts and Methods," in *Introduction to Optimum Design (Fourth Edition)*, J. S. Arora, Ed., Boston: Academic Press, 2017, pp. 771–794, ISBN: 978-0-12-800806-5. DOI: https://doi.org/10.1016/B978-0-12-800806-5.00018-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128008065000184.

[51] M. Demirci and P. Bettinger, "Using mixed integer multi-objective goal programming for stand tending block designation: A case study from Turkey," *Forest Policy and Economics*, vol. 55, pp. 28–36, Jun. 2015, ISSN: 13899341. DOI: 10.1016/j.forpol.2015.03.007.

[52] C. Hamarat, J. H. Kwakkel, E. Pruyt, and E. T. Loonen, "An exploratory approach for adaptive policymaking by using multi-objective robust optimization," *Simulation Modelling Practice and Theory*, vol. 46, pp. 25–39, Aug. 2014, ISSN: 1569190X. DOI: 10.1016/j.simpat.2014.02.008.

[53] E. C. (EC), "Europe 2020: a strategy for smart, sustainable and inclusive growth," *Working paper* {*COM (2010) 2020*}, 2010.

[54] M. Haasnoot, J. H. Kwakkel, W. E. Walker, and J. ter Maat, "Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world," *Global Environmental Change*, vol. 23, no. 2, pp. 485–498, Apr. 2013, ISSN: 09593780. DOI: 10.1016/j.gloenvcha.2012.12.006.

[55] J. C. J. Kwadijk *et al.*, "Using adaptation tipping points to prepare for climate change and sea level rise: a case study in the Netherlands," *WIREs Climate Change*, vol. 1, no. 5, pp. 729–740, Sep. 2010, ISSN: 1757-7780. DOI: https://doi.org/10.1002/wcc.64. [Online]. Available: https://doi.org/10.1002/wcc.64.

[56] A. Offermans, M. Haasnoot, and P. Valkering, "A method to explore social response for sustainable water management strategies under changing conditions," *Sustainable development*, vol. 19, no. 5, pp. 312–324, 2011, ISSN: 0968-0802.

[57] J. H. Kwakkel, M. Haasnoot, and W. E. Walker, "Developing dynamic adaptive policy pathways: a computer-assisted approach for developing adaptive strategies for a deeply uncertain world," *Climatic Change*, vol. 132, no. 3, pp. 373–386, Oct. 2015, ISSN: 0165-0009. DOI: 10.1007/s10584-014-1210-4. [Online]. Available: http://link.springer.com/10.1007/s10584-014-1210-4.

[58] Y. Gao, W. Wang, J. Shi, and N. Yu, "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5357–5369, 2020, ISSN: 1949-3053.

[59] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *Journal of Intelligent & Robotic Systems*, vol. 86, no. 2, pp. 153–173, 2017, ISSN: 0921-0296.

[60] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019, ISSN: 1553-877X.

[61] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2063–2079, 2018, ISSN: 2162-237X.

[62] B. E. Nyong-Bassey *et al.*, "Reinforcement learning based adaptive power pinch analysis for energy management of stand-alone hybrid energy storage systems considering uncertainty," *Energy*, vol. 193, p. 116 622, Feb. 2020, ISSN: 03605442. DOI: 10.1016/j.energy.2019.116622.

[63] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, "Deep Reinforcement Learning for Joint Bidding and Pricing of Load Serving Entity," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6366–6375, 2019, ISSN: 1949-3061. DOI: 10.1109/TSG.2019.2903756.

[64] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, "Deep Reinforcement Learning for Strategic Bidding in Electricity Markets," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1343–1355, 2020, ISSN: 1949-3061. DOI: 10.1109/TSG.2019.2936142.

[65] X. Xu *et al.*, "Data-driven game-based pricing for sharing rooftop photovoltaic generation and energy storage in the residential building cluster under uncertainties," *IEEE Transactions on Industrial Informatics*, 2020, ISSN: 1551-3203.

[66] E. Subramanian *et al.*, "LEarn: A Reinforcement Learning Based Bidding Strategy for Generators in Single Sided Energy Markets," in *Proceedings of the Tenth ACM International Conference on Future Energy Systems*, ser. e-Energy '19, New York, NY, USA: Association for Computing Machinery, 2019, pp. 121–127, ISBN: 9781450366717. DOI: 10.1145/3307772.3328281. [Online]. Available: https://doi.org/10.1145/3307772.3328281.

[67] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," in *Learning for Dynamics and Control*, PMLR, 2020, pp. 486–489.

[68] R. Bellman, "A Markovian decision process," *Journal of mathematics and mechanics*, vol. 6, no. 5, pp. 679–684, 1957, ISSN: 0095-9057.

[69] L. Marescot *et al.*, "Complex decisions made simple: a primer on stochastic dynamic programming," *Methods in Ecology and Evolution*, vol. 4, no. 9, pp. 872–884, 2013, ISSN: 2041-210X.

[70] F. Grabski, "1 - Discrete state space Markov processes," in *Semi-Markov Processes: Applications in System Reliability and Maintenance*, F. Grabski, Ed., Elsevier, 2015, pp. 1–17, ISBN: 978-0-12-800518-7. DOI: https://doi.org/10.1016/B978-0-12-800518-7.00001-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128005187000016.

[71] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014, ISBN: 1118625870.

[72] B. Enserink, J. H. Kwakkel, and S. Veenman, "Coping with uncertainty in climate policy making: (Mis)understanding scenario studies," *Futures*, vol. 53, pp. 1–12, Sep. 2013, ISSN: 00163287. DOI: 10.1016/j.futures.2013.09.006.

[73] I. Chadès *et al.*, "Optimization methods to solve adaptive management problems," *Theoretical Ecology*, vol. 10, no. 1, pp. 1–20, 2017, ISSN: 1874-1746. DOI: 10.1007/s12080-016-0313-0. [Online]. Available: https://doi.org/10.1007/s12080-016-0313-0.

[74] I. Chades, J. Carwardine, T. Martin, S. Nicol, R. Sabbadin, and O. Buffet, "MOMDPs: A Solution for Modelling Adaptive Management Problems," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, Jul. 2012. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/8171.

[75] C. J. Walters and R. Hilborn, "Adaptive control of fishing systems," *Journal of the Fisheries Board of Canada*, vol. 33, no. 1, pp. 145–159, 1976, ISSN: 0706-652X.

[76] D. M. Southwell, C. E. Hauser, and M. A. McCarthy, "Learning about colonization when managing metapopulations under an adaptive management framework," *Ecological Applications*, vol. 26, no. 1, pp. 279–294, 2016, ISSN: 1051-0761.

[77] S. Nicol, J. Brazill-Boast, E. Gorrod, A. McSorley, N. Peyrard, and I. Chadès, "Quantifying the impact of uncertainty on threat management for biodiversity," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019, ISSN: 2041-1723.

[78] B. K. Williams and F. A. Johnson, "Frequencies of decision making and monitoring in adaptive resource management," *PLoS One*, vol. 12, no. 8, e0182934, 2017, ISSN: 1932-6203.

[79] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive Power System Emergency Control Using Deep Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1171–1182, 2020, ISSN: 1949-3061. DOI: 10.1109/TSG.2019.2933191.

[80] H. Farhangi, "The path of the smart grid," *IEEE power and energy magazine*, vol. 8, no. 1, pp. 18–28, 2009, ISSN: 1540-7977.

[81] H. Li, Z. Wan, and H. He, "Real-Time Residential Demand Response," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4144–4154, 2020, ISSN: 1949-3061. DOI: `10.1109/TSG.2020.2978061`.

[82] E. Mocanu *et al.*, "On-Line Building Energy Optimization Using Deep Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698–3708, 2019, ISSN: 1949-3061. DOI: `10.1109/TSG.2018.2834219`.

[83] B. V. Mbuwir, F. Spiessens, and G. Deconinck, "Self-learning agent for battery energy management in a residential microgrid," in *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 2018, pp. 1–6. DOI: `10.1109/ISGTEurope.2018.8571568`.

[84] G. Zhang *et al.*, "A data-driven approach for designing STATCOM additional damping controller for wind farms," *International Journal of Electrical Power and Energy Systems*, vol. 117, p. 105 620, May 2020, ISSN: 01420615. DOI: `10.1016/j.ijepes.2019.105620`.

[85] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002, ISSN: 1941-0026. DOI: `10.1109/4235.996017`.

[86] H. Ishibuchi, N. Tsukamoto, Y. Hitotsuyanagi, and Y. Nojima, "Effectiveness of Scalability Improvement Attempts on the Performance of NSGA-II for Many-Objective Problems," in *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '08, New York, NY, USA: Association for Computing Machinery, 2008, pp. 649–656, ISBN: 9781605581309. DOI: `10.1145/1389095.1389225`. [Online]. Available: `https://doi.org/10.1145/1389095.1389225`.

[87] G. Van Rossum and F. L. Drake Jr, *Python reference manual.* Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[88] J. H. Kwakkel, "The Exploratory Modeling Workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making," *Environmental Modelling and Software*, vol. 96, pp. 239–250, Oct. 2017, ISSN: 13648152. DOI: `10.1016/j.envsoft.2017.06.054`.

[89] C. Hamarat, J. H. Kwakkel, and E. Pruyt, "Adaptive Robust Design under deep uncertainty," *Technological Forecasting and Social Change*, vol. 80, no. 3, pp. 408–418, Mar. 2013, ISSN: 00401625. DOI: `10.1016/j.techfore.2012.10.004`.

[90] S. Eker and J. H. Kwakkel, "Including robustness considerations in the search phase of Many-Objective Robust Decision Making," *Environmental Modelling and Software*, vol. 105, pp. 201–216, Jul. 2018, ISSN: 13648152. DOI: `10.1016/j.envsoft.2018.03.029`.

[91] Austrlian Energy Market Operator, "The National Electricity Market (NEM) operates on one of the world's longest interconnected power systems, stretching from Port Douglas in Queensland to The National Electricity Market FACT SHEET," Tech. Rep., 2020. [Online]. Available: `https://aemo.com.au/-/media/files/electricity/nem/national-electricity-market-fact-sheet.pdf`.

[92] A. M. Rojas-Arevalo, "Sustainability transitions modelling and assessment of socio-technical energy systems: An australian case," In Submission., Ph.D. dissertation, The University of Melbourne, 2022.

[93] A. Rojas and F. J. de Haan, "Socio-technical representation of electricity provision across scales," in *Modelling Transitions*, Routledge, 2019, pp. 139–161, ISBN: 0429056575.

[94] A. T. Crooks and A. J. Heppenstall, "Introduction to agent-based modelling," in *Agent-based models of geographical systems*, Springer, 2012, pp. 85–105.

[95] L. Ji, B. Zhang, G. Huang, Y. Cai, and J. Yin, "Robust regional low-carbon electricity system planning with energy-water nexus under uncertainties and complex policy guidelines," *Journal of Cleaner Production*, vol. 252, p. 119 800, Apr. 2020, ISSN: 09596526. DOI: 10.1016/j.jclepro.2019.119800.

[96] T. McDaniels, T. Mills, R. Gregory, and D. Ohlson, "Using expert judgments to explore robust alternatives for forest management under climate change," *Risk Analysis: An International Journal*, vol. 32, no. 12, pp. 2098–2112, 2012, ISSN: 0272-4332.

[97] E. A. Moallemi, L. Aye, J. M. Webb, F. J. de Haan, and B. A. George, *India's on-grid solar power development: Historical transitions, present status and future driving forces*, Mar. 2017. DOI: 10.1016/j.rser.2016.11.032.

[98] G. Brockman *et al.*, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[99] M. Cullen, B. Davey, K. J. Friston, and R. J. Moran, "Active inference in OpenAI Gym: A paradigm for computational investigations into psychiatric illness," *Biological psychiatry: cognitive neuroscience and neuroimaging*, vol. 3, no. 9, pp. 809–818, 2018, ISSN: 2451-9022.

[100] E. Liang *et al.*, "RLlib: Abstractions for Distributed Reinforcement Learning," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Mar. 2018, pp. 3053–3062. [Online]. Available: http://proceedings.mlr.press/v80/liang18b.html.

[101] J. H. Kwakkel, M. Haasnoot, and W. E. Walker, "Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for model-based decision support under deep uncertainty," *Environmental Modelling and Software*, vol. 86, pp. 168–183, Dec. 2016, ISSN: 13648152. DOI: 10.1016/j.envsoft.2016.09.017.

[102] J. H. Kwakkel and E. Pruyt, "Exploratory Modeling and Analysis, an approach for model-based foresight under deep uncertainty," *Technological Forecasting and Social Change*, vol. 80, no. 3, pp. 419–431, Mar. 2013, ISSN: 00401625. DOI: 10.1016/j.techfore.2012.10.005.

[103] K. Arnold, J. Gosling, and D. Holmes, *The Java programming language.* Addison Wesley Professional, 2005, ISBN: 0321349806.

[104] F. Ueckerdt, L. Hirth, G. Luderer, and O. Edenhofer, "System LCOE: What are the costs of variable renewables?" *Energy*, vol. 63, pp. 61–75, Dec. 2013, ISSN: 03605442. DOI: 10.1016/j.energy.2013.10.072.

[105] "2020 Integrated System Plan For the National Electricity Market," Australian Energy Market Operator, Tech. Rep., 2020. [Online]. Available: https://aemo.com.au/-/media/files/major-publications/isp/2020/final-2020-integrated-system-plan.pdf.

[106] "Policy Options For Australia's Electricity Supply Sector," Climate Change Authority, Tech. Rep., Aug. 2016. [Online]. Available: https://www.climatechangeauthority.gov.au/sites/default/files/2020-06/SR%20Electricity%20research%20report/Electricity%20research%20report%20-%20for%20publication.pdf.

[107] P. Nidras, W. Gerardi, P. Galanis, and R. Gawler, "Modelling illustrative electricity sector emissions reduction policies," Jacobs Group (Australia) Pty Limited, Tech. Rep., 2017. [Online]. Available: https://www.climatechangeauthority.gov.au/sites/default/files/2020-07/170217_jacobs_final_reportrevised.pdf.

[108] A. Kemp, D. Young, T. Chen, and S. Arthur, "Peer review of electricity modelling for the Climate Change Authority," Houston Kemp Economists, Tech. Rep., 2016. [Online]. Available: https://www.climatechangeauthority.gov.au/sites/default/files/2020-06/SR%20Modelling%20reports/Peer%20review%20report.pdf.

[109] T. S. Brinsmead, J. Hayward, P. Graham, J. Hayward, and P. Graham, "Australian electricity market analysis report to 2020 and 2030," CSIRO, Tech. Rep., 2014. [Online]. Available: https://arena.gov.au/assets/2017/02/CSIRO-Electricity-market-analysis-for-IGEG.pdf.

[110] "Victorian Annual Planning Report - Electricity transmission planning for Victoria," Australian Energy Market Operator, Tech. Rep., 2020. [Online]. Available: https://aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/vapr/2020/2020-vapr.pd.

[111] T. Wood and J. Ha, "Go for net zero," Grattam Institute, Tech. Rep., Apr. 2021. [Online]. Available: www.grattan.edu.au/donate..

[112] "Renewable Energy Action Plan," State of Victoria Department of Environment, Land, Water and Planning, Tech. Rep., 2017. [Online]. Available: https://www.energy.vic.gov.au/__data/assets/pdf_file/0027/74088/REAP-FA5-web.pdf.

[113] "Victoria's Climate Change Adaptation Plan (2017-2020)," State of Victoria Department of Environment, Land, Water and Planning, Tech. Rep., 2016. [Online]. Available: https://www.climatechange.vic.gov.au/__data/assets/pdf_file/0024/60729/Victorias-Climate-Change-Adaptation-Plan-2017-2020.pdf.

[114] H. R. Maier, S. Razavi, Z. Kapelan, L. S. Matott, J. Kasprzyk, and B. A. Tolson, *Introductory overview: Optimization using evolutionary algorithms and other metaheuristics*, Apr. 2019. DOI: 10.1016/j.envsoft.2018.11.018.

[115] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis, *Reinforcement Learning, Fast and Slow*, May 2019. DOI: 10.1016/j.tics.2019.02.006.

[116] S. Meng, M. Siriwardana, and J. McNeill, "The environmental and economic impact of the carbon tax in Australia," *Environmental and Resource Economics*, vol. 54, no. 3, pp. 313–332, 2013, ISSN: 0924-6460.

[117] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005, ISSN: 0936-577X.

[118] A. Saltelli, "Sensitivity Analysis for Importance Assessment," *Risk Analysis*, vol. 22, no. 3, pp. 579–590, Jun. 2002, ISSN: 0272-4332. DOI: https://doi.org/10.1111/0272-4332.00040. [Online]. Available: https://doi.org/10.1111/0272-4332.00040.

[119] Y. Gan *et al.*, "A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model," *Environmental Modelling & Software*, vol. 51, pp. 269–285, Jan. 2014, ISSN: 1364-8152. DOI: 10.1016/J.ENVSOFT.2013.09.031.

[120] M. Tosin, A. M. A. Côrtes, and A. Cunha, "A Tutorial on Sobol' Global Sensitivity Analysis Applied to Biological Models," in *Networks in Systems Biology: Applications for Disease Modeling*, F. A. B. da Silva, N. Carels, M. Trindade dos Santos, and F. J. P. Lopes, Eds., Cham: Springer International Publishing, 2020, pp. 93–118, ISBN: 978-3-030-51862-2. DOI: 10.1007/978-3-030-51862-2{_}6. [Online]. Available: https://doi.org/10.1007/978-3-030-51862-2_6.

[121] X.-Y. Zhang, M. Trame, L. Lesko, and S. Schmidt, "Sobol Sensitivity Analysis: A Tool to Guide the Development and Evaluation of Systems Pharmacology Models," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 4, no. 2, pp. 69–79, Feb. 2015, ISSN: 2163-8306. DOI: 10.1002/PSP4.6. [Online]. Available: https://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/psp4.6%20https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/psp4.6%20https://ascpt.onlinelibrary.wiley.com/doi/10.1002/psp4.6.

[122] A. Saltelli *et al.*, *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008, ISBN: 0470725176.

[123] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, no. 1-3, pp. 271–280, Feb. 2001, ISSN: 0378-4754. DOI: 10.1016/S0378-4754(00)00270-6.

[124] R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly, "Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory," *The Journal of chemical physics*, vol. 59, no. 8, pp. 3873–3878, 1973, ISSN: 0021-9606.

[125] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," *Technometrics*, vol. 33, no. 2, pp. 161–174, 1991, ISSN: 0040-1706.

[126] J. Nossent, P. Elsen, and W. Bauwens, "Sobol' sensitivity analysis of a complex environmental model," *Environmental Modelling & Software*, vol. 26, no. 12, pp. 1515–1525, Dec. 2011, ISSN: 1364-8152. DOI: 10.1016/J.ENVSOFT.2011.08.010.

[127] J. Herman and W. Usher, "SALib: An open-source Python library for Sensitivity Analysis," *The Journal of Open Source Software*, vol. 2, no. 9, Jan. 2017. DOI: 10.21105/joss.00097. [Online]. Available: https://doi.org/10.21105/joss.00097.

[128] M. D. McKay, R. J. Beckman, and W. J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979, ISSN: 00401706. DOI: 10.2307/1268522. [Online]. Available: http://www.jstor.org/stable/1268522.

[129] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices," *Computer physics communications*, vol. 145, no. 2, pp. 280–297, 2002, ISSN: 0010-4655.

[130] antonia-had, "antonia-had/SA_verification: first script for blogpost," Sep. 2020. DOI: 10.5281/ZENODO.4030955. [Online]. Available: https://zenodo.org/record/4030955.

[131] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi medical journal*, vol. 24, no. 3, pp. 69–71, 2012, ISSN: 1995-7262.

[132] S. Bankes, "The use of complexity for policy exploration," *The SAGE handbook of complexity and management*, pp. 570–589, 2011.

[133] P. M. Reed, D. Hadka, J. D. Herman, J. R. Kasprzyk, and J. B. Kollat, "Evolutionary multiobjective optimization in water resources: The past, present, and future," *Advances in Water Resources*, vol. 51, pp. 438–456, Jan. 2013, ISSN: 03091708. DOI: 10.1016/j.advwatres.2012.01.005.

[134] J. B. Kollat and P. M. Reed, “The Value of Online Adaptive Search: A Performance Comparison of NSGAII, ϵ-NSGAII and ϵMOEA,” in *Evolutionary Multi-Criterion Optimization*, C. A. Coello Coello, A. Hernández Aguirre, and E. Zitzler, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 386–398, ISBN: 978-3-540-31880-4.

[135] P. A. Vikhar, “Evolutionary algorithms: A critical review and its future prospects,” in *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, 2016, pp. 261–265. DOI: `10.1109/ICGTSPICC.2016.7955308`.

[136] M. Giuliani and A. Castelletti, “Is robustness really robust? How different definitions of robustness impact decision-making under climate change,” *Climatic Change 2016 135:3*, vol. 135, no. 3, pp. 409–424, Jan. 2016, ISSN: 1573-1480. DOI: `10.1007/S10584-015-1586-9`. [Online]. Available: `https://link.springer.com/article/10.1007/s10584-015-1586-9`.

[137] J. H. Kwakkel, S. Eker, and E. Pruyt, “How Robust is a Robust Policy? Comparing Alternative Robustness Metrics for Robust Decision-Making,” *International Series in Operations Research and Management Science*, vol. 241, pp. 221–237, 2016. DOI: `10.1007/978-3-319-33121-8{\_}10`. [Online]. Available: `https://link.springer.com/chapter/10.1007/978-3-319-33121-8_10`.

[138] M. K. Starr, *Product design and decision theory.* Prentice-Hall, 1963.

[139] L. J. Savage, “The Theory of Statistical Decision,” *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 55–67, Mar. 1951, ISSN: 0162-1459. DOI: `10.1080/01621459.1951.10500768`. [Online]. Available: `https://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500768`.

[140] A. Wald, “Statistical decision functions.,” 1950.

[141] E. Bartholomew and J. H. Kwakkel, “On considering robustness in the search phase of Robust Decision Making: A comparison of Many-Objective Robust Decision Making, multi-scenario Many-Objective Robust Decision Making, and Many Objective Robust Optimization,” *Environmental Modelling and Software*, vol. 127, p. 104 699, May 2020, ISSN: 13648152. DOI: `10.1016/j.envsoft.2020.104699`.

[142] E. Zitzler and L. Thiele, “Multiobjective optimization using evolutionary algorithms—a comparative case study,” in *International conference on parallel problem solving from nature*, Springer, 1998, pp. 292–301.

[143] K. Deb and R. B. Agrawal, “Simulated binary crossover for continuous search space,” *Complex systems*, vol. 9, no. 2, pp. 115–148, 1995, ISSN: 0891-2513.

[144] K. Deb, “Multi-objective optimisation using evolutionary algorithms: an introduction,” in *Multi-objective evolutionary optimisation for product design and manufacturing*, Springer, 2011, pp. 3–34.

[145] C. J. C. H. Watkins and P. Dayan, “Q-learning,” *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992, ISSN: 0885-6125.

[146] V. Mnih *et al.*, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.

[147] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[148] M. Glavic, *(Deep) Reinforcement learning for electric power system control and related problems: A short review and perspectives*, Jan. 2019. DOI: `10.1016/j.arcontrol.2019.09.008`.

[149] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum learning for reinforcement learning domains: A framework and survey," *arXiv preprint arXiv:2003.04960*, 2020.

[150] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey.," *Journal of Machine Learning Research*, vol. 10, no. 7, 2009, ISSN: 1532-4435.

[151] M. C. Stanton and K. Roelich, "Decision making under deep uncertainties: A review of the applicability of methods in practice," *Technological Forecasting and Social Change*, vol. 171, p. 120 939, Oct. 2021, ISSN: 0040-1625. DOI: `10.1016/J.TECHFORE.2021.120939`.

[152] T. Homma and A. Saltelli, "Importance measures in global sensitivity analysis of nonlinear models," *Reliability Engineering & System Safety*, vol. 52, no. 1, pp. 1–17, Apr. 1996, ISSN: 0951-8320. DOI: `10.1016/0951-8320(96)00002-6`.

[153] F. Klinker, "Exponential moving average versus moving exponential average," *Mathematische Semesterberichte*, vol. 58, no. 1, pp. 97–107, 2011, ISSN: 1432-1815. DOI: `10.1007/s00591-010-0080-8`. [Online]. Available: `https://doi.org/10.1007/s00591-010-0080-8`.

Appendices

Appendix A

GSE Appendix

A.1 GSE Uncertain Parameters

Table A.1: List of 33 uncertain parameters for the *GSE*. One parameter from Rojas-Arevalo's original study, *onsiteGeneration*, was omitted from this study as it restricted the *GSE* from loading previously generated (scenario specific) forecast data from an existing database, adding a considerable amount of runtime per *GSE* execution.

Name	Description	Value Range (BAU)
annualCpi	Adjusts future tariffs to 2019 values. Past tariffs are adjusted with a conversion table for the consumer price (CPI) index quered from the database.	1 to 5% (2.33%)
annualInflation	Impacts only the prices of electricity offered by generators - i.e. the base price or LCOE.	1 to 5% (3.3%)
consumption	Market operator's consumption forecast.	Central, Fast, High DET, Slow, Step (Central)
domesticConsumptionPercentage	Percentage of residential consumption in Victoria.	20 to 50% (30%)
energyEfficiency	Market operator's energy efficiency forecast.	Central, Slow, Step (Central)

Continued on next page

Table A.1 – continued from previous page

Name	Description	Value Range (BAU)
generationRolloutPeriod	Time to roll out new generation. The nameplate capacity is divided by the number of years to represent an incremental deployment.	1 to 10 (1)
generatorRetirement	Shift in years on closure date of brown coal power plants.	-5 to 5 (0)
importPriceFactor	Premium paid for imported electricity. Applied to the wholesale price when local demand is unmet.	-50 to 50% (29%)
includePublicallyAnnouncedGen	Decision variable to include emerging projects. These are projects that are not yet completely approved. New generators, their status and potential dates of operation are obtained from the "Generation Information" released by AEMO each month.	True or False (False)
learningCurve	Decreases the base price of wind and solar generators.	0 to 10% (5%)
nameplateCapacityChangeBattery	Changes the nameplate capacity of battery generators.	-50 to 50% (0%)
nameplateCapacityChangeBrownCoal	Changes the nameplate capacity of brown coal generators.	-50 to 50% (0%)

Continued on next page

Table A.1 – continued from previous page

Name	Description	Value Range (BAU)
nameplateCapacityChangeCcgt	Changes the nameplate capacity of combined cycle gas turbine generators.	-50 to 50% (0%)
nameplateCapacityChangeOcgt	Changes the nameplate capacity of open cycle gas turbine generators.	-50 to 50% (0%)
nameplateCapacityChangeSolar	Changes the nameplate capacity of solar generators.	-50 to 50% (0%)
nameplateCapacityChangeWater	Changes the nameplate capacity of hydro generators.	-50 to 50% (0%)
nameplateCapacityChangeWind	Changes the nameplate capacity of wind generators.	-50 to 50% (0%)
nonScheduleGenSpotMarket	Sets the market in which non-scheduled generation can participate. Only generation with a minimum capacity defined by *nonScheduleMinCapMarketGen* can be included in the market selected.	Primary, Secondary, None (None)
nonScheduleMinCapMarketGen	Sets the minimum nameplate capacity in MW required for non-scheduled generation to participate in a market.	0.1 to 30MW (30MW)
priceChangePercentageBattery	Changes the BAU electricity base price for batteries.	-50 to 50% (0%)
priceChangePercentageBrownCoal	Changes the BAU electricity base price for brown coal generators.	-50 to 50% (0%)

Continued on next page

Table A.1 – continued from previous page

Name	Description	Value Range (BAU)
priceChangePercentageCcgt	Changes the BAU electricity base price for combined cycle gas turbine generators.	-50 to 50% (0%)
priceChangePercentageOcgt	Changes the BAU electricity base price for open cycle gas turbine generators.	-50 to 50% (0%)
priceChangePercentageSolar	Changes the BAU electricity base price for solar generators.	-50 to 50% (0%)
priceChangePercentageWater	Changes the BAU electricity base price for hydro generators.	-50 to 50% (0%)
priceChangePercentageWind	Changes the BAU electricity base price for wind generators.	-50 to 50% (0%)
rooftopPV	Changes the ISP forecast on uptake of rooftop PB in residential, business or both sectors.	Business, Residential, Both (Both)
scheduleMinCapMarketGen	Sets the minimum nameplate capacity in MW required for schedule generation to participate in the market.	10 to 30MW (30MW)
semiScheduleGenSpotMarket	Sets the market in which semi-scheduled generation can participate. Only generation with a minimum capacity defined by *semiScheduleMinCapMarketGen* can be included in the market selected.	Primary, Secondary, None (None)

Continued on next page

Table A.1 – continued from previous page

Name	Description	Value Range (BAU)
semiScheduleMinCapMarketGen	Sets the minimum nameplate capacity in MW required for semi-schedule generation to participate in a market.	0.1 to 30MW (30MW)
solarUptake	Market operator's solar uptake forecast.	Central, Slow, Step (Central)
technologicalImprovement	Increases the capacity factors of wind and solar generators by adding - not compounding - a constant factor every year.	0 to 10% (5%)
wholesaleTariffContribution	Varies the percentage contributions from wholesale prices to the final electricity tariff. Upper and lowers bounds derived by Rojas-Arevalo [92] via analysis of historic yearly average contributions of wholesale prices from the primary spot market in Victoria.	10 to 45% (28.37%)

A.2 GSE Output Values

Table A.2: *GSE* output variables. Variables with an asterisk are not included in the state space of the RL agents in chapter 8.

Name	Description
Avg Tariff per household ($¢/KWh$)	Annual average tariff price per household.
Consumption per household (KWh)	Annual electricity consumption per household.
GHG Emissions per household (tCO_2e)	Annual average GHG emissions per household
Number of Active Actors	Number of generators participating in the market.
Number of Domestic Consumers (households)	Number of households whose electricity is supplied by the market.
Percentage Renewable Production (%)	Annual average market share of renewable electricity.
Primary Max Unmet Demand Per Hour (MWh)*	Annual maximum quantity of electricity demand that was unmet in a given hour in the primary market.
Primary Total Unmet Demand (Days)	Annual number of days where electricity demand was unmet in the primary market.
Primary Total Unmet Demand (Hours)*	Annual number of hours where electricity demand was unmet in the primary market.
Primary Total Unmet Demand (MWh)	Annual quantity of electricity demand that was unmet in the primary market.
Primary Wholesale ($\$/MWh$)	Annual average primary wholesale electricity price.
Secondary Max Unmet Demand Per Hour (MWh)*	Annual maximum quantity of electricity demand that was unmet in a given hour in the secondary market.

Continued on next page

Table A.2 – continued from previous page

Name	Description
Secondary Total Unmet Demand (Days)	Annual number of days where electricity demand was unmet in the secondary market.
Secondary Total Unmet Demand (Hours)*	Annual number of hours where electricity demand was unmet in the secondary market.
Secondary Total Unmet Demand (MWh)	Annual quantity of electricity demand that was unmet in the secondary market.
System Production Battery (MWh)*	Annual quantity of electricity supplied by coal-powered generators.
System Production Coal (MWh)	Annual quantity of electricity supplied by coal-powered generators.
System Production Gas (MWh)	Annual quantity of electricity supplied by gas-powered generators.
System Production Off Spot (MWh)	Annual quantity of electricity supplied outside of the spot markets.
System Production Primary Spot (MWh)	Annual quantity of electricity supplied in the primary spot market.
System Production Rooftop PV (MWh)*	Annual quantity of electricity supplied by rooftop PV cells.
System Production Secondary Spot (MWh)	Annual quantity of electricity supplied in the secondary spot market.
System Production Solar (MWh)	Annual quantity of electricity supplied by solar-powered generators.
System Production Water (MWh)	Annual quantity of electricity supplied by water-powered generators.
System Production Wind (MWh)	Annual quantity of electricity supplied by wind-powered generators.

Continued on next page

Table A.2 – continued from previous page

Name	Description
Year	Simulated year for results.

Appendix B

Model Preparation Appendix

B.1 Forecast Data Smoothing Algorithms

Algorithm 2 Smooth forecast electricity demand for different bidding window sizes.

```
    **Input:** $D = [d_0, \ldots, d_{47}]$, 30 minute electricity demands for calendar day.
    **Input:** $h$, hours between bidding rounds.
    **Output:** $S$, array of smoothed forecast electricity demands, s.t. $|S| = 24/h$.
procedure SmoothDemand($D, h$)
    $t \leftarrow 0$
    $i \leftarrow 0$
    $w \leftarrow 2 \cdot h$                              ▷ 30min intervals between bids
    while $t < 48$ do                                    ▷ Loop over day
        $S[i] \leftarrow \sum_{j=t}^{t+w-1} D[j]$
        $t \leftarrow t + w$
    return $S$                                ▷ Smoothed electricity demands
```

Algorithm 3 Smooth forecast solar capacity for different bidding window sizes.

```
    **Input:** $C = [c_0, \ldots, c_{47}]$, 30 minute solar capacities for calendar day.
    **Input:** $h$, hours between bidding rounds.
    **Output:** $S$, array of smoothed forecast solar capacities, s.t. $|S| = 24/h$.
procedure SmoothSolarCapacity($D, h$)
    $t \leftarrow 0$
    $i \leftarrow 0$
    $w \leftarrow 2 \cdot h$                              ▷ 30min intervals between bids
    while $t < 48$ do                                    ▷ Loop over day
        $S[i] \leftarrow \sum_{j=t}^{t+w-1} C[j]/w$
        $t \leftarrow t + w$
    return $S$                                   ▷ Smoothed solar capacities
```

B.2 Historical Validation Results

Table B.1: Historical validation results for *GHGE Levels*. Units are in tCO_2e.

Year	Historical	0.5hr	1hr	2hr	3hr	4hr	6hr	8hr	12hr	24hr
1998	59.20	50.45	50.45	50.45	50.45	50.45	50.45	50.45	50.45	50.45
1999	61.80	56.50	57.05	57.05	57.05	57.05	57.05	57.05	57.05	57.05
2000	62.70	59.14	59.94	59.96	59.96	59.96	59.96	59.96	59.96	59.97
2001	62.20	59.13	60.00	59.99	59.99	59.99	59.99	59.99	59.99	60.00
2002	60.40	59.36	60.26	60.25	60.25	60.25	60.25	60.25	60.25	60.25
2003	61.50	60.49	61.34	61.38	61.38	61.38	61.38	61.38	61.38	61.38
2004	64.70	61.30	62.04	62.05	62.05	62.05	62.05	62.05	62.05	62.05
2005	63.50	63.35	63.19	63.20	63.20	63.20	63.20	63.20	63.20	63.20
2006	64.30	65.46	64.74	65.11	65.11	65.11	65.11	65.11	65.11	65.11
2007	63.30	65.46	64.74	64.90	64.90	64.90	64.90	64.90	64.90	64.91
2008	63.70	65.40	64.65	65.40	65.40	65.40	65.40	65.40	65.41	65.41
2009	65.60	65.27	64.54	65.22	65.22	65.22	65.22	65.21	65.21	65.21
2010	65.30	65.11	64.43	65.41	65.41	65.41	65.41	65.41	65.41	65.40
2011	64.40	63.99	63.74	64.45	64.45	64.45	64.44	64.44	64.43	64.40
2012	66.70	62.79	63.31	63.86	63.85	63.85	63.84	63.82	63.80	63.74
2013	59.00	61.50	62.64	63.13	63.12	63.11	63.09	63.07	63.02	62.89
2014	57.00	61.07	62.17	63.00	62.99	62.98	62.95	62.93	62.89	62.75
2015	61.10	59.06	59.58	60.77	60.76	60.75	60.72	60.69	60.64	60.48
2016	59.20	59.88	60.49	61.72	61.71	61.70	61.68	61.65	61.61	61.47
2017	56.10	48.62	48.44	49.15	49.13	49.12	49.08	49.05	48.98	48.78
2018	46.40	46.25	46.08	46.78	46.77	46.76	46.74	46.71	46.67	46.53

Table B.2: Historical validation results for *Renewable Market Share*. Units are in %.

Year	**Historical**	**0.5hr**	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
2005	5.33	6.95	7.54	7.53	7.53	7.53	7.53	7.53	7.53	7.52
2006	5.03	8.02	9.22	8.77	8.77	8.77	8.77	8.77	8.77	8.76
2007	2.92	7.99	9.24	9.04	9.04	9.03	9.03	9.03	9.03	9.02
2008	3.15	7.91	9.28	8.35	8.35	8.35	8.35	8.35	8.35	8.34
2009	5.21	8.45	9.86	9.12	9.12	9.12	9.12	9.12	9.12	9.11
2010	8.30	9.49	11.09	9.95	9.95	9.95	9.96	9.96	9.96	9.98
2011	6.70	9.76	11.02	10.22	10.22	10.23	10.24	10.25	10.27	10.33
2012	8.21	9.42	9.37	8.66	8.67	8.68	8.70	8.71	8.75	8.86
2013	12.03	10.06	8.88	8.23	8.24	8.26	8.29	8.33	8.40	8.61
2014	9.61	11.27	10.21	9.10	9.12	9.14	9.17	9.21	9.28	9.51
2015	12.05	13.56	13.42	11.89	11.91	11.93	11.98	12.02	12.11	12.39
2016	15.53	14.16	13.87	12.32	12.34	12.36	12.40	12.44	12.52	12.75
2017	14.72	21.13	21.55	20.56	20.61	20.65	20.73	20.82	20.99	21.50
2018	18.98	23.42	23.73	22.72	22.76	22.79	22.86	22.93	23.07	23.49
2019	21.60	24.32	24.33	22.92	22.94	22.96	22.99	23.02	23.08	23.28
2020	23.02	24.01	22.12	22.66	22.67	22.69	22.72	22.75	22.81	22.99

Table B.3: Historical validation results for *Wholesale Prices*. Units are in *$/MWh*.

Year	**Historical**	**0.5hr**	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
2005	29.82	58.66	84.27	84.17	84.06	83.94	83.72	83.49	83.04	81.72
2006	36.83	52.61	87.79	86.54	86.42	86.31	86.09	85.86	85.42	84.14
2007	69.02	45.85	65.63	64.70	64.61	64.52	64.34	64.16	63.80	62.76
2008	42.57	43.02	54.91	52.66	52.59	52.52	52.38	52.24	51.97	51.18
2009	41.71	39.41	49.31	44.93	44.87	44.80	44.67	44.55	44.29	43.55
2010	38.94	37.43	42.56	40.89	40.82	40.75	40.61	40.48	40.01	39.24
2011	30.67	33.95	31.12	29.13	29.04	28.95	28.76	28.58	28.13	27.09
2012	45.80	26.08	23.10	23.39	23.31	23.23	23.08	22.92	22.59	21.68
2013	53.58	24.85	20.21	19.89	19.81	19.73	19.58	19.43	19.11	17.77
2014	42.86	33.47	22.92	22.21	22.10	21.98	21.73	21.49	21.01	19.53
2015	34.10	43.78	33.16	31.71	31.49	31.27	30.82	30.38	29.49	26.83
2016	49.35	40.25	31.59	31.32	31.14	30.95	30.57	30.19	29.40	26.89
2017	94.25	57.08	69.62	67.78	67.48	67.15	67.17	66.43	65.01	59.83
2018	95.68	66.88	81.36	80.52	80.01	79.48	79.19	78.42	77.52	71.89
2019	120.01	119.06	103.80	103.44	102.58	101.70	99.99	98.35	95.45	84.13
2020	70.53	64.24	77.96	77.92	77.39	76.85	75.63	74.53	72.46	65.04

Table B.4: Historical validation results for *Tariff Prices*. Units are in ¢/*KWh*.

Year	**Historical**	**0.5hr**	**1hr**	**2hr**	**3hr**	**4hr**	**6hr**	**8hr**	**12hr**	**24hr**
2001	15.70	7.71	7.50	7.48	7.47	7.44	7.42	7.36	7.21	27.32
2002	10.39	6.87	6.89	6.88	6.86	6.84	6.82	6.77	6.65	27.03
2003	18.99	11.49	11.51	11.49	11.47	11.43	11.40	11.33	11.12	N/A
2004	20.72	14.85	14.71	14.69	14.66	14.62	14.58	14.50	14.25	N/A
2005	20.24	18.44	18.45	18.42	18.39	18.34	18.29	18.19	17.88	N/A
2006	24.92	36.75	36.70	36.65	36.60	36.50	36.40	36.21	35.63	N/A
2007	14.39	22.10	22.00	21.98	21.95	21.89	21.84	21.73	21.42	23.38
2008	11.58	14.45	14.45	14.43	14.41	14.37	14.34	14.26	14.05	25.55
2009	21.22	27.67	26.78	26.75	26.71	26.64	26.57	26.44	26.03	27.74
2010	19.93	24.25	20.83	20.80	20.78	20.72	20.66	20.55	20.21	24.96
2011	24.29	29.09	27.21	27.16	27.11	27.02	26.92	26.73	26.16	25.77
2012	31.34	30.57	27.36	27.26	27.15	26.94	26.74	26.20	24.99	29.10
2013	13.66	10.09	10.12	10.08	10.04	9.96	9.89	9.73	9.26	31.92
2014	12.13	8.75	8.72	8.70	8.67	8.62	8.58	8.48	8.20	30.89
2015	27.14	17.26	16.19	16.09	16.00	15.80	15.60	15.20	13.96	27.81
2016	34.96	25.23	24.82	24.65	24.48	24.13	23.79	23.10	21.05	27.82
2017	24.35	17.25	17.14	17.05	16.95	16.74	16.53	16.08	14.61	29.44
2018	23.99	29.27	29.23	29.11	28.97	29.14	28.79	28.10	25.42	23.31
2019	22.31	23.84	23.85	23.73	23.61	23.50	23.31	23.33	21.79	21.77

Appendix C

Augmenting Uncertainty Appendix

C.1 Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) [128] is a method for pseudo-randomly sets of values from multi-dimensional spaces. It is widely utilised in deep uncertainty literature [141], as it ensures that the values of each dimension are represented evenly across the entire sampled set [47]. In this study, the multi-dimensional space is the space of all possible scenarios for the *GSE*, where each dimension is one of the 33 uncertain parameters of the *GSE* (Appendix A.1). LHS requires specifying the number of samples (scenarios) required by the user. Given a specified number of scenarios, s, LHS divides the range of each uncertain parameter into s equal bins, and samples one value from each bin, resulting in s distinct values. To construct a scenario, a value is selected from the s sampled values for each uncertain parameter. This scenario construction is repeated until all s scenarios have been created. All sampling of scenarios in this paper is completed using LHS.

C.2 Sobol Indices

Sobol Indices [123] is a model independent, GSA technique, that is based on decomposing the variance of a given output value of concern. To determine the Sobol Indices of a model's n different uncertain parameters, a probabilistic perspective is considered on the model's input parameters, such that the input is a random vector $U \in \mathbb{R}^n$ represented as:

$$X = M(U) = M(U_1, \dots, U_n)$$

where $X \in \mathbb{R}$ is the model output of concern.

For a set of input parameters, Sobol suggested to decompose the function M into summands with increasing dimensionality:

$$X = M_0 + \sum_{i=1}^{n} M_i(U_i) + \sum_{1 \leq i < j \leq n} M_{ij}(U_i, U_j) + \cdots + M_{1..n}(U_i, \dots, U_n)$$

where:
$M_0 = \mathbb{E}[X]$
$M_i(U_i) = \mathbb{E}[X|U_i] - M_0$
$M_{ij}(U_i, U_j) = \mathbb{E}[X|U_i, U_j] - M_0 - M_i - M_j$
$\dots$

such that M_0 is the expected value of X, and values of increasing order are recursively defined conditional expected values [120]. From this, the total variance in the output value can be represented as:

$$Var(X) = \sum_{k} Var(M_k(U_k)), \qquad \text{s.t.} \quad \varnothing \neq k \subset \{1, \dots, n\}$$

where $Var(M_k(U_k))$ is the conditional variance of U_k, and contains the uncertain parameters who are sampled according to the subset k. Therefore, the Sobol Index of the subset of uncertain parameters denoted by k is the ratio between the contribution to the output value by the uncertain parameters within k, and the total variance of the output value:

$$S_k = \frac{Var(M_k(Y_k))}{Var(X)}$$

Following this equation, for $k \subset \{1, \dots, n\}, k \neq \emptyset$, the sum of all Sobol Indices are equal to 1.

$$\sum_{k} S_k = \sum_{i=1}^{n} S_i + \sum_{1 \leq i < j \leq n} S_{ij} + \cdots + S_{1..n} = 1$$

The order of a Sobol Index determines the insights into the impacts and interactions of uncertain parameters in the model the index can provide [120]. First order indices measure the contribution of U_i to the total variance of output value X [126]. Alternatively, it can be considered as the (expected) fraction of the output value's variance that would be removed if U_i were to be fixed [126]. First order indices are defined as:

$$S_i = \frac{Var(M_i(U_i))}{X}, i = 1, \ldots, n$$

Second order indices add an additional uncertain parameter, U_j, and calculate the contribution and interactions of U_i and U_j:

$$S_{ij} = \frac{Var(M_{ij}(U_ij))}{Var(X)}, 1 \leq i < j \leq n$$

Such construction of the Sobol Indices can be done using more uncertain parameters, to identify the different contributions and interactions between parameters. In addition, a value known as the total Sobol Index [152] for an uncertain parameter can be computed. Total Sobol Indices S^T compute the full contribution of U_i to the total variance, considering all orders of Sobol Indices.

$$S_i^T = \sum_{\substack{k \subset \{1,\ldots,n\} \\ i \in k}} S_k \qquad i = 1, \ldots, n$$

C.3 Sobol Indices Results

Table C.1: Sobol Indices - *GHGE Levels*.

	S1		ST	
Uncertain Parameter	**Median**	**Max**	**Median**	**Max**
domesticConsumptionPercentage	0.431463	0.643033	0.465691	0.662504
nameplateCapacityChangeBrownCoal	0.160818	0.318758	0.222259	0.353898
consumption	0.043654	0.100362	0.067349	0.150037
priceChangePercentageBrownCoal	0.041143	0.109466	0.110609	0.217251
priceChangePercentageWind	0.016998	0.065526	0.069585	0.126808
nameplateCapacityChangeWind	0.014701	0.035224	0.052023	0.060377
generatorRetirement	0.011720	0.848524	0.021551	0.932369
priceChangePercentageWater	0.010176	0.031380	0.034546	0.068517
nameplateCapacityChangeWater	0.001673	0.006412	0.012510	0.013258
learningCurve	0.000854	0.008169	0.015394	0.036049
nameplateCapacityChangeSolar	0.000773	0.003466	0.005183	0.008689
nonScheduleMinCapMarketGen	0.000461	0.000832	0.000215	0.000739
priceChangePercentageCcgt	0.000219	0.000801	0.000186	0.000550
nonScheduleGenSpotMarket	0.000211	0.001612	0.000584	0.001701
includePublicallyAnnouncedGen	0.000206	0.000068	0.000080	0.000503
semiScheduleGenSpotMarket	0.000202	0.000041	0.000158	0.000516
scheduleMinCapMarketGen	0.000199	0.000019	0.000221	0.000526
importPriceFactor	0.000164	0.000629	0.000081	0.000550
nameplateCapacityChangeOcgt	0.000136	0.003313	0.000139	0.003845
rooftopPV	0.000131	0.000694	0.001623	0.004772
annualCpi	0.000106	0.000942	0.000073	0.000530
priceChangePercentageSolar	0.000079	0.000668	0.000110	0.000553
generationRolloutPeriod	0.000078	0.007252	0.016452	0.030729
semiScheduleMinCapMarketGen	0.000070	0.004374	0.009572	0.018025
technologicalImprovement	0.000058	0.000418	0.000113	0.000561
energyEfficiency	0.000054	0.000923	0.000211	0.000752
priceChangePercentageOcgt	0.000046	0.000614	0.000076	0.000535
annualInflation	0.000039	0.000971	0.000152	0.000653
priceChangePercentageBattery	0.000038	0.000391	0.000195	0.000730
solarUptake	0.000031	0.000988	0.001716	0.004140
nameplateCapacityChangeCcgt	0.000023	0.000639	0.000201	0.000507
nameplateCapacityChangeBattery	0.000023	0.000925	0.000116	0.000579
wholesaleTariffContribution	0.000009	0.000405	0.000193	0.000725

Table C.2: Sobol Indices - *Renewable Market Share*.

	S1		**ST**	
Uncertain Parameter	**Median**	**Max**	**Median**	**Max**
nameplateCapacityChangeBrownCoal	0.300210	0.785678	0.403402	0.873972
priceChangePercentageBrownCoal	0.116118	0.235053	0.281622	0.422706
nameplateCapacityChangeWind	0.043432	0.136078	0.106786	0.190687
priceChangePercentageWind	0.042516	0.127547	0.181685	0.251495
priceChangePercentageWater	0.025920	0.069706	0.087901	0.153378
generatorRetirement	0.021296	0.958307	0.036776	0.973439
consumption	0.017770	0.001858	0.038763	0.163584
nameplateCapacityChangeWater	0.007145	0.035239	0.028130	0.054407
learningCurve	0.005602	0.023136	0.043468	0.069374
rooftopPV	0.000488	0.000680	0.002552	0.008926
nameplateCapacityChangeCcgt	0.000465	0.000989	0.000405	0.002120
semiScheduleGenSpotMarket	0.000450	0.001047	0.000359	0.002114
priceChangePercentageSolar	0.000409	0.001214	0.000224	0.002094
domesticConsumptionPercentage	0.000366	0.000299	0.000414	0.002111
priceChangePercentageOcgt	0.000349	0.001264	0.000196	0.002300
importPriceFactor	0.000322	0.000292	0.000197	0.002293
priceChangePercentageBattery	0.000277	0.000313	0.000413	0.002094
scheduleMinCapMarketGen	0.000264	0.000229	0.000468	0.002132
includePublicallyAnnouncedGen	0.000256	0.000003	0.000200	0.002087
annualCpi	0.000234	0.001508	0.000207	0.002214
semiScheduleMinCapMarketGen	0.000219	0.000779	0.000240	0.002237
technologicalImprovement	0.000211	0.000429	0.000256	0.002248
solarUptake	0.000133	0.002083	0.002644	0.006967
wholesaleTariffContribution	0.000124	0.000457	0.000390	0.002238
energyEfficiency	0.000121	0.000237	0.000449	0.002114
nameplateCapacityChangeOcgt	0.000113	0.023649	0.001382	0.025742
nonScheduleGenSpotMarket	0.000074	0.001791	0.001012	0.002980
nameplateCapacityChangeBattery	0.000056	0.000728	0.000216	0.002268
nameplateCapacityChangeSolar	0.000053	0.002920	0.007602	0.015943
priceChangePercentageCcgt	0.000007	0.001150	0.000357	0.002277
nonScheduleMinCapMarketGen	0.000004	0.000934	0.000428	0.002207
generationRolloutPeriod	2.80×10^{-7}	0.013919	0.030006	0.072395
annualInflation	1.31×10^{-7}	0.002739	0.000278	0.002267

Table C.3: Sobol Indices - *Wholesale Prices*.

	S1		**ST**	
Uncertain Parameter	**Median**	**Max**	**Median**	**Max**
consumption	0.057733	0.186906	0.237841	0.570887
nameplateCapacityChangeBrownCoal	0.055466	0.677626	0.449189	0.836375
importPriceFactor	0.036200	0.071432	0.085651	0.135825
nameplateCapacityChangeWind	0.028912	0.119166	0.237297	0.606252
annualInflation	0.019931	0.063219	0.041057	0.115406
generationRolloutPeriod	0.016797	0.189495	0.181518	0.742689
generatorRetirement	0.013760	0.156183	0.126162	0.545121
nonScheduleGenSpotMarket	0.012657	0.081262	0.053675	0.339448
priceChangePercentageOcgt	0.005330	0.030499	0.014375	0.044014
nameplateCapacityChangeWater	0.005053	0.020434	0.059608	0.109959
scheduleMinCapMarketGen	0.003274	0.018614	0.078897	0.264053
nameplateCapacityChangeSolar	0.003254	0.013686	0.021550	0.035302
nameplateCapacityChangeOcgt	0.002825	0.021705	0.028074	0.110454
priceChangePercentageCcgt	0.001180	0.005606	0.002888	0.008254
priceChangePercentageBrownCoal	0.000887	0.029749	0.002490	0.033989
nameplateCapacityChangeBattery	0.000879	0.017924	0.099476	0.283960
rooftopPV	0.000677	0.007393	0.006577	0.023680
solarUptake	0.000416	0.004689	0.005750	0.031444
priceChangePercentageBattery	0.000336	0.033534	0.046182	0.191041
nameplateCapacityChangeCcgt	0.000191	0.002661	0.000515	0.006362
priceChangePercentageWater	0.000153	0.001909	0.000624	0.006788
learningCurve	0.000122	0.003615	0.000944	0.006581
wholesaleTariffContribution	0.000079	0.001676	0.000221	0.006444
semiScheduleGenSpotMarket	0.000074	0.001793	0.000220	0.006446
priceChangePercentageSolar	0.000042	0.002471	0.000214	0.006458
annualCpi	0.000032	0.004723	0.000293	0.006908
includePublicallyAnnouncedGen	0.000029	0.002171	0.000223	0.006680
domesticConsumptionPercentage	0.000025	0.002462	0.000269	0.006393
technologicalImprovement	0.000024	0.002846	0.000377	0.006868
energyEfficiency	0.000019	0.003670	0.000188	0.006240
semiScheduleMinCapMarketGen	0.000009	0.002027	0.000271	0.006492
nonScheduleMinCapMarketGen	0.000007	0.001575	0.000294	0.006562
priceChangePercentageWind	0.000001	0.009579	0.002526	0.018020

Table C.4: Sobol Indices - *Tariff Prices*.

	S1		**ST**	
Uncertain Parameter	**Median**	**Max**	**Median**	**Max**
wholesaleTariffContribution	0.097860	0.345687	0.238982	0.452541
nameplateCapacityChangeBrownCoal	0.042759	0.744182	0.426918	0.873710
consumption	0.027486	0.128416	0.200478	0.555375
generationRolloutPeriod	0.024005	0.963153	0.177147	0.982268
nameplateCapacityChangeWind	0.022471	0.089508	0.168062	0.538636
importPriceFactor	0.016905	0.040269	0.062917	0.115676
annualCpi	0.015158	0.070229	0.035442	0.132687
generatorRetirement	0.014246	0.133075	0.104980	0.465400
annualInflation	0.011083	0.030658	0.024162	0.096082
nonScheduleGenSpotMarket	0.007451	0.044419	0.031451	0.244640
priceChangePercentageOcgt	0.005053	0.011647	0.014894	0.025679
nameplateCapacityChangeOcgt	0.004142	0.016702	0.019049	0.124381
priceChangePercentageBattery	0.003386	0.026684	0.044553	0.186332
nameplateCapacityChangeBattery	0.002267	0.018350	0.084584	0.270836
nameplateCapacityChangeWater	0.001961	0.029085	0.046588	0.133640
nameplateCapacityChangeSolar	0.001565	0.009612	0.011775	0.017680
priceChangePercentageCcgt	0.001547	0.005310	0.004974	0.014277
priceChangePercentageWind	0.001069	0.013254	0.002842	0.022981
semiScheduleMinCapMarketGen	0.000937	0.006864	0.003905	0.017319
rooftopPV	0.000875	0.003895	0.004399	0.014687
domesticConsumptionPercentage	0.000606	0.002697	0.001288	0.016380
solarUptake	0.000564	0.003712	0.004265	0.015816
learningCurve	0.000411	0.005644	0.001443	0.015504
semiScheduleGenSpotMarket	0.000385	0.016962	0.002444	0.027532
priceChangePercentageBrownCoal	0.000373	0.022543	0.002937	0.033463
energyEfficiency	0.000359	0.004101	0.001459	0.015085
scheduleMinCapMarketGen	0.000331	0.015167	0.067067	0.336423
nameplateCapacityChangeCcgt	0.000207	0.004225	0.001721	0.016560
priceChangePercentageWater	0.000096	0.003812	0.001589	0.015734
technologicalImprovement	0.000058	0.003060	0.001649	0.015117
includePublicallyAnnouncedGen	0.000024	0.002239	0.001092	0.014326
priceChangePercentageSolar	0.000014	0.004148	0.001279	0.016042
nonScheduleMinCapMarketGen	0.000010	0.004428	0.001328	0.016236

Table C.5: Sobol Indices - *Unmet Demand Days*.

	S1		**ST**	
Uncertain Parameter	**Median**	**Max**	**Median**	**Max**
nameplateCapacityChangeBrownCoal	0.251168	0.815854	0.551708	0.906850
consumption	0.131273	0.371806	0.354537	0.593065
nameplateCapacityChangeWind	0.110605	0.172955	0.273726	0.573481
nameplateCapacityChangeOcgt	0.010744	0.025341	0.041535	0.120860
nameplateCapacityChangeWater	0.010017	0.033263	0.055669	0.160904
nameplateCapacityChangeSolar	0.008458	0.021746	0.027402	0.042458
solarUptake	0.003054	0.015850	0.007374	0.045531
rooftopPV	0.002941	0.013204	0.009798	0.093998
generatorRetirement	0.002151	0.271412	0.049243	0.812973
nameplateCapacityChangeBattery	0.000182	0.001741	0.000059	0.001720
technologicalImprovement	0.000105	0.001285	0.000026	0.001801
semiScheduleGenSpotMarket	0.000094	0.001425	0.000023	0.001780
energyEfficiency	0.000010	0.002409	0.000002	0.001628
scheduleMinCapMarketGen	0.000010	0.000504	0.000003	0.001679
priceChangePercentageOcgt	0.000009	0.002049	0.000002	0.001785
learningCurve	0.000008	0.001980	0.000002	0.001779
domesticConsumptionPercentage	0.000006	0.001922	0.000002	0.001685
annualInflation	0.000004	0.002276	0.000002	0.001698
priceChangePercentageWind	0.000003	0.001315	0.000002	0.001736
wholesaleTariffContribution	0.000003	0.000461	0.000002	0.001602
nonScheduleMinCapMarketGen	0.000003	0.002178	0.000002	0.001695
nonScheduleGenSpotMarket	0.000002	0.000406	0.000003	0.001836
generationRolloutPeriod	0.000002	0.017681	0.000004	0.144904
nameplateCapacityChangeCcgt	0.000002	0.003550	0.000037	0.001709
semiScheduleMinCapMarketGen	0.000001	0.001545	0.000003	0.001799
includePublicallyAnnouncedGen	0.000001	0.001518	0.000002	0.001801
annualCpi	0.000001	0.002911	0.000002	0.001708
importPriceFactor	0.000001	0.000910	0.000002	0.001744
priceChangePercentageCcgt	0.000001	0.002432	0.000002	0.001781
priceChangePercentageBrownCoal	0.000001	0.002720	0.000002	0.001771
priceChangePercentageBattery	3.84×10^{-7}	0.001534	0.000002	0.001628
priceChangePercentageWater	2.52×10^{-7}	0.000369	0.000002	0.001712
priceChangePercentageSolar	6.35×10^{-8}	0.001975	0.000002	0.001529

C.4 Ordered Uncertain Parameters

Table C.6: Pearson correlations coefficients computed during the factor fixing algorithm in chapter 6.

Top Uncertainties Sampled	**GHGE Levels** (tCO_2e)	**Renewable Market Share** $(\%)$	**Wholesale Prices** $(\$/MWh)$	**Tariff Prices** $(¢/KWh)$	**Unmet Demand Days** $(Days)$
1	0.483	0.367	0.183	0.412	0.336
2	0.535	0.485	0.458	0.509	0.555
3	0.582	0.514	0.482	0.537	0.686
4	0.810	0.565	0.513	0.550	0.732
5	0.812	0.581	0.532	0.550	0.770
6	0.845	0.592	0.545	0.642	0.799
7	0.865	0.742	0.690	0.748	0.832
8	0.925	0.756	0.731	0.827	0.855
9	0.965	0.797	0.735	0.831	0.931
10	0.968	0.825	0.740	0.839	0.949
11	0.974	0.842	0.774	0.861	0.954
12	0.977	0.889	0.854	0.887	0.969
13	0.985	0.913	0.898	0.920	0.970
14	0.990	0.948	0.919	0.932	0.973
15	0.992	0.949	0.946	0.995	0.978
16	1.000	1.000	1.000	1.000	1.000

Table C.7: Uncertain parameters ordered by their maximum median S1 value using the Sobol Indices results (Appendix C.3) for all five performance indicators.

Rank	Uncertain Parameter	Median S1 (Max)
1	domesticConsumptionPercentage	0.431463
2	nameplateCapacityChangeBrownCoal	0.300210
3	consumption	0.131273
4	priceChangePercentageBrownCoal	0.116118
5	nameplateCapacityChangeWind	0.110605
6	wholesaleTariffContribution	0.097860
7	priceChangePercentageWind	0.042516
8	importPriceFactor	0.036200
9	priceChangePercentageWater	0.025920
10	generationRolloutPeriod	0.024005
11	generatorRetirement	0.021296
12	annualInflation	0.019931
13	annualCpi	0.015158
14	nonScheduleGenSpotMarket	0.012657
15	nameplateCapacityChangeOcgt	0.010744
16	nameplateCapacityChangeWater	0.010017
17	nameplateCapacityChangeSolar	0.008458
18	learningCurve	0.005602
19	priceChangePercentageOcgt	0.005330
20	priceChangePercentageBattery	0.003386
21	scheduleMinCapMarketGen	0.003274
22	solarUptake	0.003054
23	rooftopPV	0.002941
24	nameplateCapacityChangeBattery	0.002267
25	priceChangePercentageCcgt	0.001547
26	semiScheduleMinCapMarketGen	0.000881
27	nameplateCapacityChangeCcgt	0.000465
28	nonScheduleMinCapMarketGen	0.000461
29	semiScheduleGenSpotMarket	0.000445
30	priceChangePercentageSolar	0.000409
31	energyEfficiency	0.000359
32	includePublicallyAnnouncedGen	0.000256
33	technologicalImprovement	0.000211

Table C.8: Final 12 uncertain parameters, ordered by their importance.

Rank	Uncertain Parameter
1	domesticConsumptionPercentage
2	nameplateCapacityChangeBrownCoal
3	consumption
4	priceChangePercentageBrownCoal
5	nameplateCapacityChangeWind
6	wholesaleTariffContribution
7	priceChangePercentageWind
8	importPriceFactor
9	priceChangePercentageWater
10	generationRolloutPeriod
11	generatorRetirement
12	annualCpi

C.5 Factor Fixing Algorithm

Algorithm 4 Factor Fixing

```
    Input: $U$, set denoting the ordering of uncertain parameters according to table C.7.
    Input: $n_s$, number of scenarios to evaluate.
    Input: $S_{\text{BAU}}$, BAU scenario values for each uncertain parameter.
    Output: $C$, Pearson correlation coefficient values.
procedure FACTOR_FIXING($U, n_s, s_{\text{BAU}}$)
    $S \leftarrow sample\_scenarios(n_s)$                ▷ $|S| = n_s$, $\forall s \in S$ ordered by $U$
    $O_d \leftarrow run\_experiments(S)$
    $C \leftarrow \{\}$                                  ▷ Correlation results, $|C| = |U| - 1$

    for $i = 1..|U|$ do                                  ▷ Loop all uncertainties
        $S_1 = \{\}$
        for $j = 0..n_s - 1$ do                          ▷ Loop all scenarios
            $s \leftarrow S[j]$
            $S_1[j] \leftarrow \{s[k] \,|0 \leq k < i\} \cup \{S_{BAU}[k] \,|i \leq k < |U|\}$
        $O_1 \leftarrow run\_experiments(S_1)$
        $C[i-1] = corr(O_d, O_1)$                        ▷ Compute Pearson correlation coefficient
    return $C$
```

Appendix D

Metro Map Appendix

D.1 Stability Tests

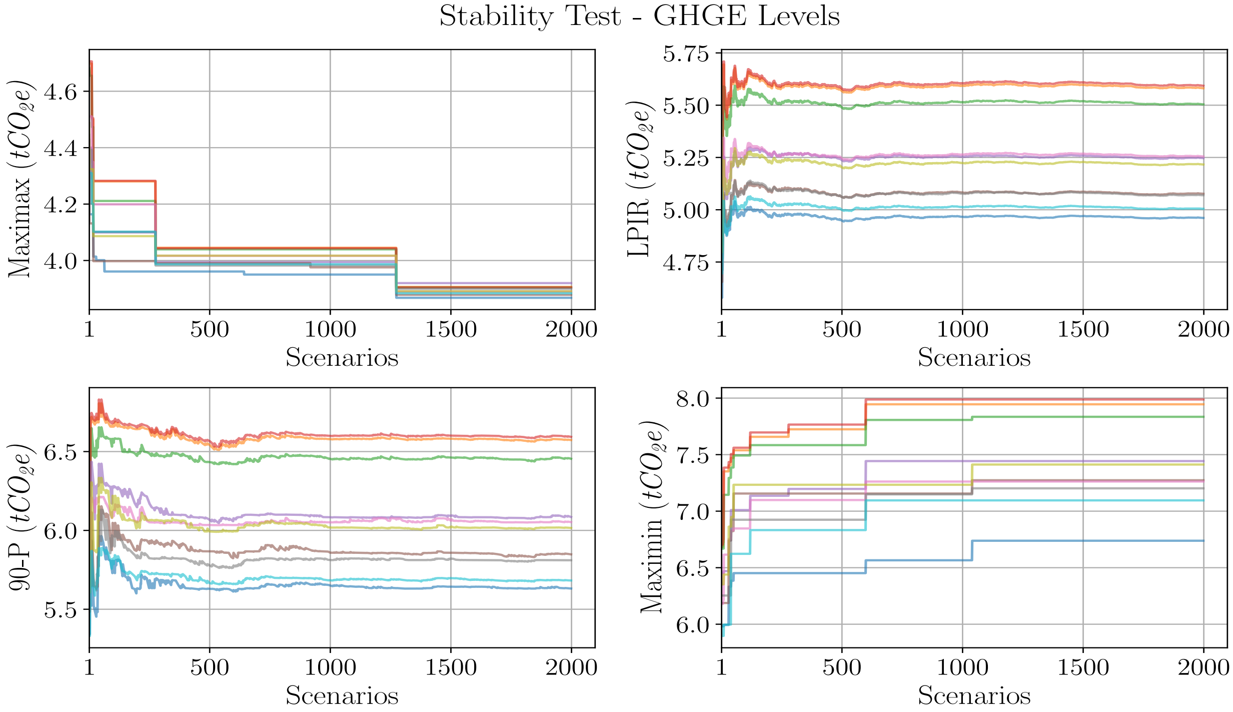


Figure D.1: Stability test results of the four robustness metrics for *GHGE Levels*.

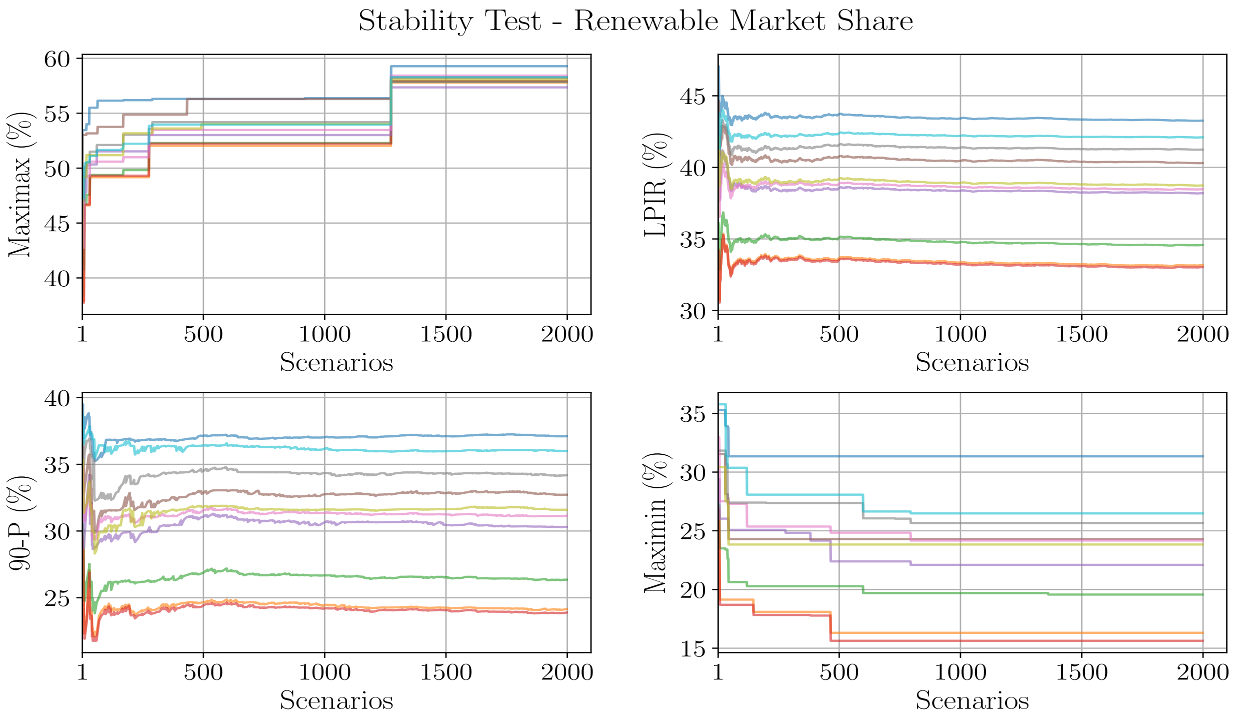


Figure D.2: Stability test results of the four robustness metrics for *Renewable Market Share*.

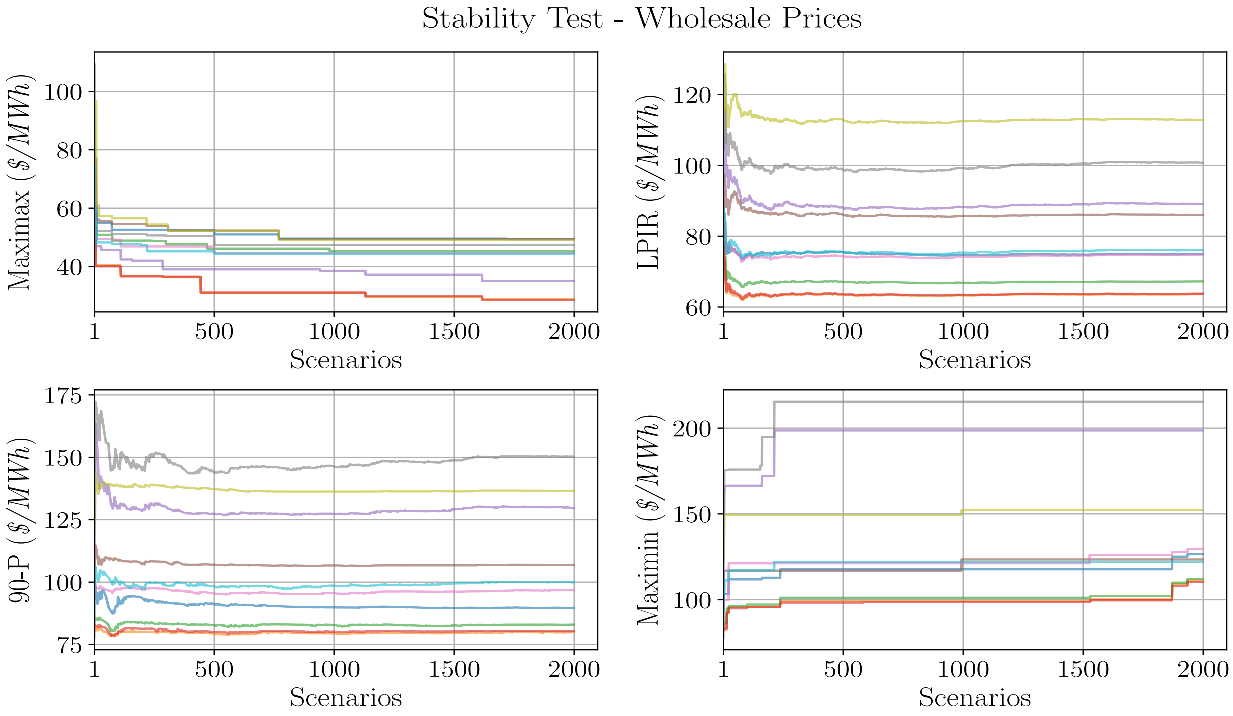


Figure D.3: Stability test results of the four robustness metrics for *Wholesale Prices*.

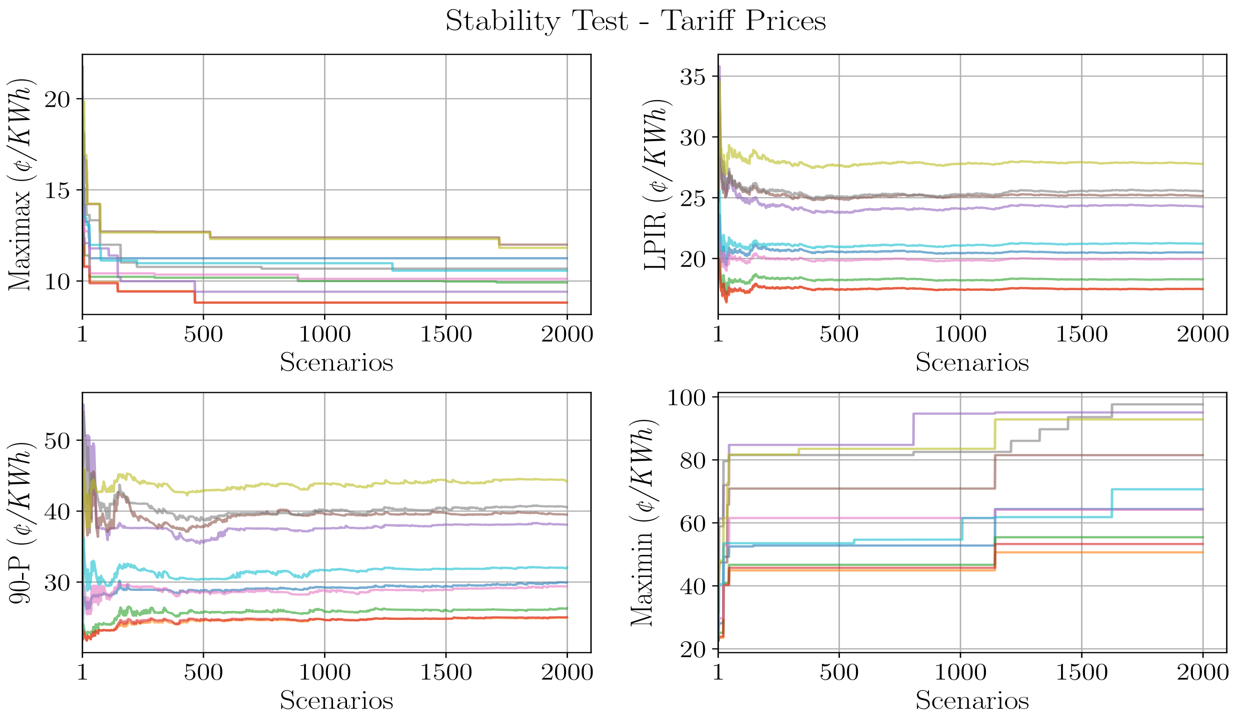


Figure D.4: Stability test results of the four robustness metrics for *Tariff Prices*.

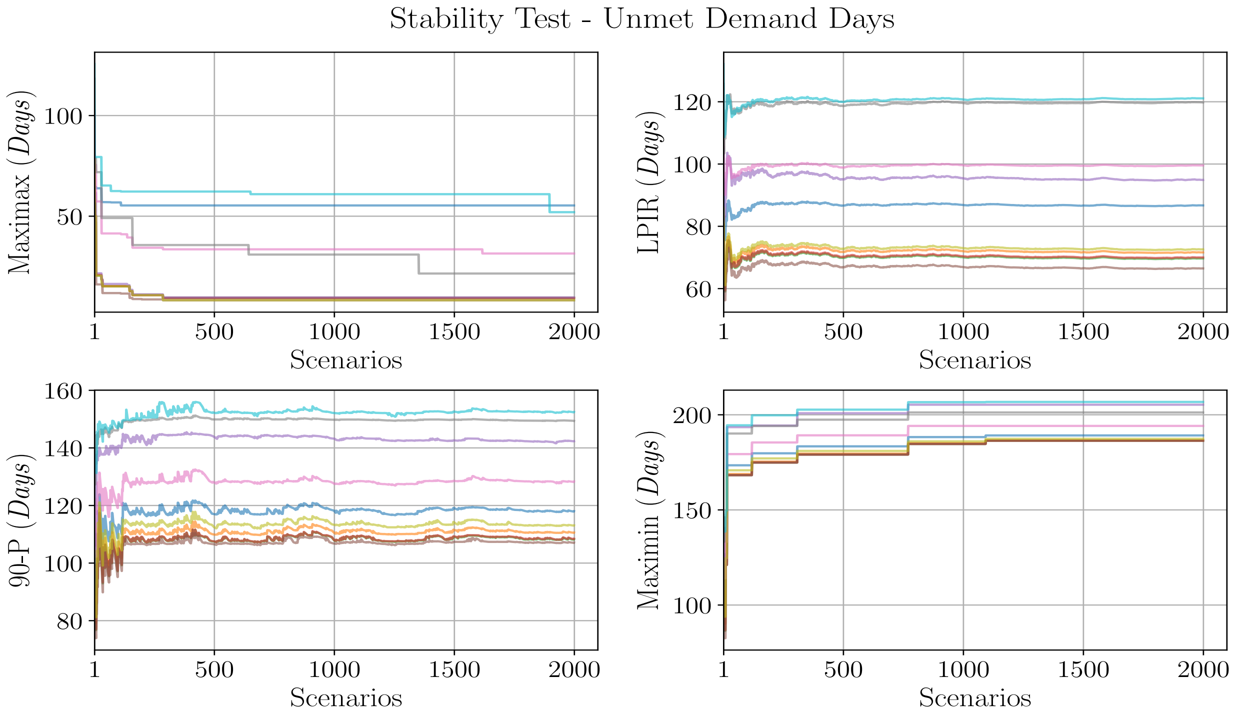


Figure D.5: Stability test results of the four robustness metrics for *Unmet Demand Days*.

D.2 ϵ-NSGAII Hyperparameters

Table D.1: Hyperparameters for the ϵ-NSGAII experiments. P_i denotes the population archive after generation i.

Hyperparameter	**Value**
Initial population	100
Cross-over probability	1.0
Mutation probability	$1/\lvert P_i\rvert$

D.3 Robustness Metrics Results

Table D.2: Robustness metric results - *GHGE Levels*. Units are in tCO_2e.

	Baseline	*Maximax*	*LPIR*	*90-P*	*Maximin*
Min	0.72	0.23	0.20	0.20	0.20
STD	1.06	0.84	0.88	0.80	0.74
Mean	2.87	1.83	1.76	1.62	1.62
IQR	1.28	1.10	1.14	1.05	0.97
25%	2.22	1.21	1.12	1.03	1.07
Median	2.75	1.70	1.61	1.49	1.51
75%	3.50	2.31	2.26	2.08	2.05
Max	6.32	6.32	6.32	6.32	6.25
Range	5.60	6.09	6.12	6.12	6.05

Table D.3: Robustness metric results - *Renewable Market Share*. Units are in *%*.

	Baseline	*Maximax*	*LPIR*	*90-P*	*Maximin*
Min	25.94	25.94	25.96	26.90	27.53
STD	11.88	10.63	11.51	10.34	9.29
Mean	51.76	67.71	68.77	71.07	71.37
IQR	17.62	14.24	15.33	14.33	12.51
25%	44.03	60.91	61.99	64.52	65.63
Median	51.56	68.75	69.99	71.74	71.86
75%	61.65	75.16	77.31	78.86	78.13
Max	81.93	90.93	91.77	91.73	91.72
Range	55.99	64.99	65.82	64.83	64.18

Table D.4: Robustness metric results - *Wholesale Prices*. Units are in *$/MWh*.

	Baseline	*Maximax*	*LPIR*	*90-P*	*Maximin*
Min	17.38	17.38	17.38	17.38	32.22
STD	22.77	37.30	30.97	29.95	29.83
Mean	81.65	115.81	109.74	111.30	110.17
IQR	35.98	45.92	38.87	35.34	37.65
25%	63.59	89.87	87.75	90.37	88.84
Median	79.88	111.27	106.45	106.61	106.88
75%	99.57	135.80	126.62	125.72	126.49
Max	130.78	324.44	219.41	219.23	306.92
Range	113.40	307.06	202.03	201.85	274.69

Table D.5: Robustness metric results - *Tariff Prices*. Units are in *¢/KWh*.

	Baseline	*Maximax*	*LPIR*	*90-P*	*Maximin*
Min	3.42	3.42	3.42	3.43	4.10
STD	9.66	15.06	14.45	14.53	13.81
Mean	15.52	24.84	23.67	24.06	23.49
IQR	10.78	17.70	16.12	16.09	15.57
25%	8.75	14.06	13.68	14.00	13.86
Median	12.75	20.91	19.85	20.33	19.97
75%	19.53	31.76	29.80	30.09	29.43
Max	52.38	128.47	125.76	126.66	126.65
Range	48.96	125.05	122.34	123.24	122.55

Table D.6: Robustness metric results - *Unmet Demand Days*. Units are in *Days*.

	Baseline	*Maximax*	*LPIR*	*90-P*	*Maximin*
Min	1.07	1.07	1.07	1.07	1.07
STD	50.57	63.93	57.45	48.54	51.25
Mean	103.39	169.57	143.00	144.06	166.33
IQR	62.68	88.77	83.41	65.64	75.84
25%	69.40	124.56	100.38	110.66	126.20
Median	99.52	171.74	140.93	142.16	164.97
75%	132.08	213.33	183.79	176.29	202.03
Max	280.17	350.48	336.31	336.45	350.48
Range	279.10	349.41	335.24	335.38	349.41

Appendix E

Reinforcement Learning Appendix

E.1 Performance Indicator Mass Experiments

Table E.1: This table displays the *GSE* results for the performance indicators. The dataset for the results was generated using 10^3 policy pathways in 10^3 scenarios (10^6 total *GSE* runs). All pathways and scenarios were randomly sampled using LHS.

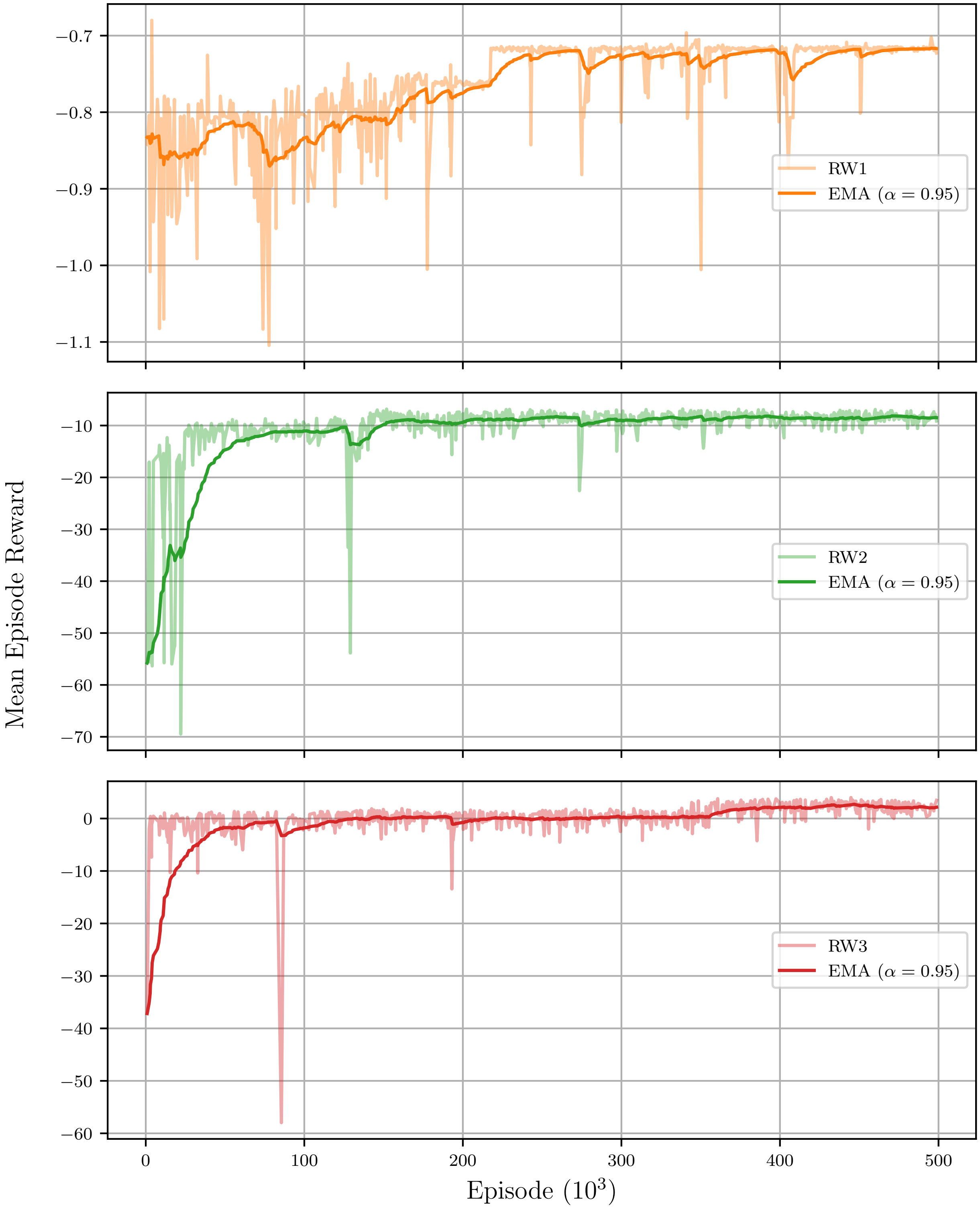
	GHGE Levels (tCO_2e)	**Renewable Market Share** ($\%$)	**Wholesale Prices** ($\$/MWh$)	**Tariff Prices** ($¢/KWh$)	**Unmet Demand Days** ($Days$)
Mean	5.25	0.38	21.93	80.46	85.68
STD	0.61	0.07	9.70	23.48	35.34
Min	0.16	0.16	1.10	23.02	8.30
25%	4.80	0.34	15.28	63.27	59.23
50%	5.15	0.38	19.23	75.51	81.32
75%	5.62	0.43	25.56	91.67	110.11
Max	8.00	0.95	107.75	213.72	358.36

E.2 Reward Function Evaluation Results

Reward Function	**Wall Time (hours)**
RF1	10.82
RF2	10.91
RF3	10.88
Total	32.61

Table E.2: Wall times for training with each reward function.

Figure E.1: Mean episode reward during training for the three reward functions evaluated for this study. Each plot has an exponential moving average [153] fitted, with $\alpha = 0.95$.

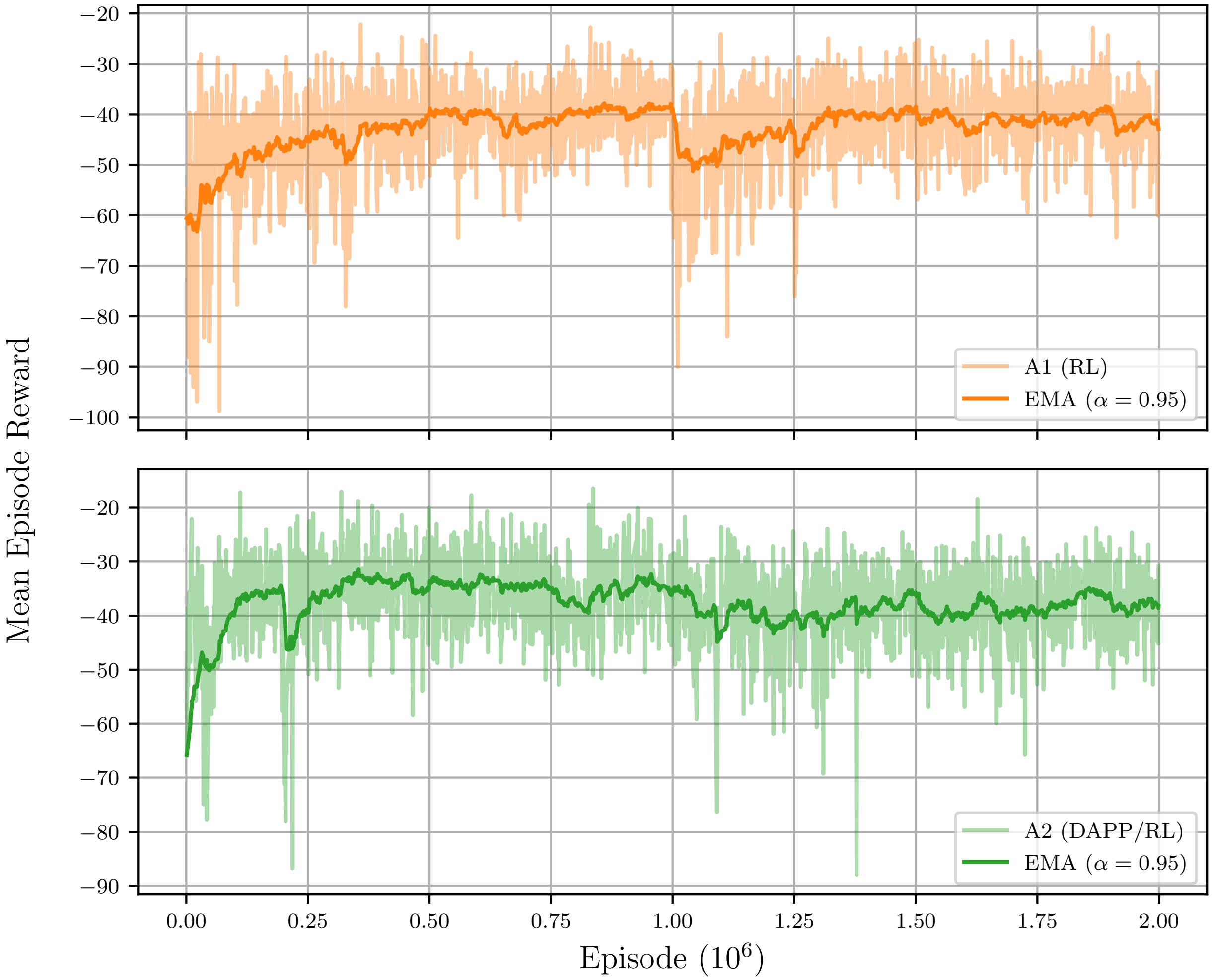


E.3 Agent Training Data

Table E.3: Runtime duration for training the two final agents, A1 and A2.

RL Agent	Wall Time (hours)
A1	47.33
A2	46.83
Total	94.16

Figure E.2: Mean episode reward during training for the two final agents, A1 and A2. Each plot has an exponential moving average [153] fitted, with $\alpha = 0.95$.



E.4 Agent Evaluation results

Table E.4: RL agent results - *GHGE Levels*. Units are in *tCO_2e*.

	A1-R	**A1**	**Baseline**	**A2**	**A2-R**
Min	0.13	0.29	0.29	0.12	0.10
STD	0.98	1.11	1.12	1.01	0.97
Mean	2.32	2.76	2.99	2.00	1.91
IQR	1.33	1.52	1.57	1.36	1.25
25%	1.59	1.93	2.14	1.25	1.20
Median	2.17	2.59	2.83	1.84	1.75
75%	2.92	3.45	3.72	2.61	2.45
Max	6.50	6.97	6.97	6.60	6.52
Range	6.37	6.68	6.68	6.48	6.43

Table E.5: RL agent results - *Renewable Market Share*. Units are in *%*.

	A1-R	**A1**	**Baseline**	**A2**	**A2-R**
Min	23.32	18.35	18.35	21.71	18.72
STD	11.35	12.74	11.83	12.90	12.11
Mean	60.04	53.56	49.74	65.58	66.39
IQR	15.63	19.01	16.88	19.15	16.62
25%	52.41	44.50	41.40	56.23	58.56
Median	60.42	54.42	49.82	66.25	67.23
75%	68.04	63.52	58.28	75.38	75.18
Max	95.09	91.06	90.85	96.01	95.28
Range	71.76	72.71	72.50	74.30	76.56

Table E.6: RL agent results - *Wholesale Prices*. Units are in *\$/MWh*.

	A1-R	**A1**	**Baseline**	**A2**	**A2-R**
Min	20.80	14.41	13.31	25.27	18.01
STD	43.94	25.28	21.95	21.39	40.09
Mean	116.18	87.90	82.90	99.28	116.78
IQR	52.42	33.83	31.08	26.02	52.81
25%	84.19	70.00	66.78	84.61	87.39
Median	104.99	85.49	81.47	96.45	104.67
75%	136.61	103.83	97.86	110.63	140.20
Max	336.75	318.20	195.62	211.45	292.26
Range	315.95	303.79	182.32	186.18	274.24

Table E.7: RL agent results - *Tariff Prices*. Units are in *¢/KWh*.

	A1-R	**A1**	**Baseline**	**A2**	**A2-R**
Min	2.30	1.67	1.38	2.35	1.82
STD	17.18	11.83	10.90	13.01	18.39
Mean	24.60	17.43	16.38	21.11	26.15
IQR	17.97	12.50	11.65	14.47	19.39
25%	12.82	9.38	8.89	11.96	13.38
Median	19.61	14.03	13.23	17.35	20.85
75%	30.78	21.88	20.55	26.43	32.76
Max	157.04	107.49	102.88	122.65	159.04
Range	154.74	105.82	101.50	120.30	157.22

Table E.8: RL agent results - *Unmet Demand Days*. Units are in *Days*.

	A1-R	**A1**	**Baseline**	**A2**	**A2-R**
Min	0.31	0.31	0.28	1.41	0.31
STD	59.80	56.09	50.84	50.96	57.21
Mean	136.82	110.77	101.94	127.99	129.93
IQR	85.18	67.16	59.71	71.87	70.69
25%	92.93	72.75	67.75	88.58	89.03
Median	129.34	102.07	95.33	117.10	114.07
75%	178.11	139.91	127.46	160.45	159.72
Max	341.41	352.72	310.28	326.52	349.00
Range	341.10	352.41	310.00	325.10	348.69