Decomposition Algorithms for Optimal Manufacturing and Power Systems Infrastructure Planning

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Department of Chemical Engineering

Cristiana L. Lara

B.S., Chemical Engineering, Federal University of Rio de Janeiro

Carnegie Mellon University
Pittsburgh, PA
May, 2019

Acknowledgments

First of all I would like to thank my advisor, Prof. Ignacio E. Grossmann, for being a true inspiration and role model. I am very grateful for his mentoring and support throughout these five years, always encouraging me to explore new areas and ideas, and sharing my excitement with these challenging projects. His passion for advancing knowledge, commitment to quantitative analysis, ethic standards, hard work and humility have shaped both my professional and personal life. I am very honored and grateful to have him as my mentor.

I am grateful to my thesis committee, Prof. Antonio Conejo, Prof. Jay Apt, Prof. Chrysanthos Gounaris, and Prof. Nick Sahinidis, for their wisdom, patience, guidance and expertise. It was an honor for me to have them in my committee.

I would like to acknowledge the financial support by Exxon Mobil Corporation; U.S. Department of Energy, Office of Fossil Energy's Crosscutting Research Program through the Institute for the Design of Advanced Energy Systems (IDAES); Wilton E. Scott Institute for Energy Innovation at Carnegie Mellon University; and the Center for Advanced Process Decision-making (CAPD).

During my time at Carnegie Mellon University I was fortunate to collaborate with many students and researchers in different projects. I thank Dr. Dharik S. Mallapragada, Dr. Dimitri J. Papageorgiou, and Dr. Aranya Venkatesh for their helpful feedback, their patience to listen and discuss my doubts and concerns, and for providing me with a great dataset that allowed me to focus on the modeling and algorithmic efforts of our project. I am also very grateful to Dr. Francisco Trespalacios for the great ideas and helpful feedbacks in our research project. I thank David E. Bernal and Can Li for all the fruitful discussions and ideas that have led to our paper together. I would also like to thank Dr. John Siirola for the patience, helpful feedback, and for helping me parallelize the SDDiP algorithm.

During these years, I also had the pleasure of spending time as a visiting researcher at PSR, National Renewable Energy Laboratory (NREL) and Sandia National Laboratory. I would like to thank everyone that I met during these visits and that spent time discussing about power systems planning models with me. A special thank to Dr. Mario Pereira for introducing the concepts of SDDP and SDDiP to me, and for all the great discussions we had, which truly impacted my research.

I want to thank the students that started the chemical engineering Ph.D. program with me in 2014. I especially thank Khal Hajj, Nick Lamson, Dr. Burcu Karagoz, and Chris Hanselman for their friendship. I also want to thank the many students and visiting researchers from PSE. The informal discussions and friendship created a relaxed but productive atmosphere at CMU. I especially thank Dr. Pablo Garcia-Herreros, Dr. Francisco Trespalacios, Dr. Qi Zhang, Dr. Rob Apap, Dr. Anirudh Subramanyam, Dr. Nikos Lappas, Qi Chen, David Bernal, Can Li, Ben Sauk and Carlos Florensa for inspiration and friendship.

I would also like to thank Prof. Ofélia Araújo and Prof. Tatiana Rappoport for encouraging me to apply to CMU's Ph.D. program. Finally, I would like to thank my husband, Matthew Orie, my parents, Cláudia and Luiz Lara, and my sister, Luana Lara, for the unconditional love and support, and for always believing in me.

Abstract

This thesis addresses two challenging problems in the area of optimal infrastructure planning: continuous centralized-distributed facility location-allocation, and power systems generation expansion planning. We focus on decomposition approaches for solving these optimization problems in face of nonconvexities, discrete decisions, integration of planning and scheduling, and optimization under uncertainty.

Part I addresses continuous facility location-allocation problems with maximum capacity, also known as Capacitated Multi-Facility Weber Problem (CMWP). This class of problems optimizes the 2-dimensional continuous location and allocation of facilities based on their maximum capacity and the given coordinates of the suppliers and customers. We propose an extended version of the classic CMWP with fixed costs, multiple types of facilities and two sets of fixed points (suppliers and consumers), and formulate it as a nonconvex Generalized Disjunctive Programming model. We propose a Bilevel Decomposition algorithm and, based on the bounding properties of the subproblems, we prove its ϵ-convergence. We benchmark our Bilevel Decomposition against commercial global optimization solvers and the results show that our method is more effective at finding global optimal solutions. We then address the design and planning of manufacturing networks considering the option of centralized and distributed facilities, which is formulated as a multi-period version of our previous CMWP. To handle this added complexity, we propose an accelerated version of the Bilevel Decomposition with additional steps to reduce the feasible space. We benchmark the performance of the Accelerated Bilevel Decomposition algorithm against the original Bilevel Decomposition and the commercial global solvers and show that the accelerated algorithm outperforms the other options. Additionally, we illustrate the applicability of the proposed model and solution framework with a case study for a biomass supply chain.

Part II addresses generation expansion planning problems in power systems. We start by investigating the impact of including operational and temporal details in a Generation Expansion Planning framework. This preliminary analysis shows that time-slice approaches tend to overestimate solar capacity and underestimate wind and natural gas capacity if compared to chronological hourly approaches. It also reveals that the latter consistently leads to lower unmet demand, implying the need for sufficient temporal resolution and chronology. Therefore, motivated by these results, we propose a deterministic Mixed-Integer Linear Programming (MILP) formulation for long-term planning of electric power infrastructure by simultaneously considering annual investment decisions and hourly operational decisions. We adopt judicious approximations and aggregations to improve its tractability and, to overcome computational challenges, we propose a decomposition algorithm based on Nested Benders Decomposition for multi-period MILP problems. Our decomposition adapts previous nested Benders methods by handling integer and continuous state variables. We apply the proposed modeling framework to a case study in the Electric Reliability Council of Texas (ERCOT) region, and demonstrate large computational savings from our decomposition. We then extend the proposed deterministic model to Multistage Stochastic Mixed-integer Programming in order to handle uncertainties. To be able to solve such large-scale model, we decompose the problem using Stochastic Dual Dynamic Integer Programming (SDDiP), and take advantage of parallel processing to solve it more efficiently. The proposed framework is also applied to a case study in the ERCOT region, and we show that the parallelized SDDiP algorithm allows the solution of instances with quadrillions of variables and constraints.

This thesis makes several contributions both in modeling and algorithms. It addresses challenging practical problems and showcases the benefits of problem-specific and structure-specific decomposition approaches to solve complex discrete optimization problems.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis addresses two challenging problems in the area of optimal infrastructure planning: continuous facility location-allocation (Part I), and power systems generation expansion planning (Part II). We focus on decomposition approaches for solving these optimization problems in face of one or more of the following complicating factors: nonconvexities, discrete decisions, integration of planning and scheduling, and optimization under uncertainty.

The detailed list of objectives for the two classes of problems is as follows:

I. Continuous facility location-allocation design and planning:

- Extend the Capacitated Multi-facility Weber Problem (CMWP), i.e., 2-dimensional continuous facility location-allocation problem with maximum capacity, to consider fixed costs, multiple types of facilities, and two sets of fixed points representing suppliers and consumers. The problem is formulated as a nonconvex Generalized Disjunctive Programming (GDP).
- Propose a Bilevel Decomposition algorithm for the extended CMWP that converges to the global optimum within an ϵ-tolerance in a finite number of iterations.
- Propose a general framework to systematically address the optimal multi-period design and planning of centralized and distributed manufacturing networks as a 2-dimensional continuous facility location-allocation problem with maximum capacity.

- Propose an accelerated version of the Bilevel Decomposition algorithm that keeps its rigor (i.e., its ϵ-convergence), but has a better performance to allow the solution of large-scale multi-period instances within a reasonable amount of time.

II. Electric power infrastructure planning:

- Propose a deterministic Mixed-integer Linear Programming (MILP) model involving multi-scale spatial and temporal strategies to address generation expansion planning with hourly operating details.
- Propose a valid Nested Decomposition algorithm to solve large-scale multi-period MILP models with mixed-integer recourse.
- Extend the deterministic formulation of the generation expansion planning with hourly operating details to Multistage Stochastic Mixed-integer Linear Programming to be able to handle strategic and operational uncertainties, and solve it efficiently with a parallel version of the Stochastic Dual Dynamic integer Programming (SDDiP) algorithm.

This chapter contains a review of the basic concepts that are dealt with in the rest of the thesis. We start by giving an overview of the applications covered in this thesis (Section 1.1): (i) Section 1.1.1 provides motivation and review of the area of centralized and distributed manufacturing networks; (ii) Section 1.1.2 provides a brief review of the research efforts in continuous facility location-allocation problems, focusing on the variant of this problem with maximum capacity; and (iii) Section 1.1.3 provides a brief literature review of contributions in the area of power systems generation expansion planning, including modeling and solution efforts for both deterministic and stochastic formulations.

Section 1.2 reviews the main modeling frameworks and solution strategies used throughout this thesis: Section 1.2.1 provides a brief overview of Generalized Disjunctive Programming (GDP), which is the modeling framework used in Part I of this thesis; Section 1.2.2 reviews fundamental concepts in Stochastic Programming (SP), which is the modeling framework used in Chapter 6; and Section 1.2.3 reviews decomposition methods for mathematical optimization, focusing on Lagrangean Decomposition, Benders Decomposition, and Nested

Benders Decomposition, which serve as basis for the solution strategies introduced in this thesis. This introductory chapter ends with a brief overview of the remaining chapters of the thesis (Section 1.3).

1.1 Applications in optimal infrastructure planning

1.1.1 Centralized and distributed manufacturing networks

Advances in technology have led to the rethinking of traditional manufacturing. In the past few decades, public and private initiatives have been sponsoring research on smaller-scale and cleaner manufacturing processes. The F3 Factory Project was launched in 2009 to enhance the competitiveness of the European chemical industry by promoting modular continuous plants with small and medium scale production (Bieringer et al., 2013). Likewise, the U.S. Advanced Manufacturing National Program Office (AMNPO) has brought together corporations, federal agencies, and universities to advance manufacturing technologies by investing in areas such as High Efficiency Modular Chemical Processes (HEMCP), Additive Manufacturing (3D printing), and Process Intensification (Ozokwelu, 2014).

Modular plants consist of manufacturing sites with their major equipment pieces in standardized modules instead of having customized site-specific design (Chen and Grossmann, 2019). Their potential advantages include higher flexibility, faster time-to-market, and improved safety (Roy, 2017). This concept is not new (Rogers and Bottaci, 1997), but combined with distributed manufacturing and the recent advances in process intensification (Moulijn et al., 2008), it can be a viable and beneficial alternative to traditional large-scale manufacturing.

The concept of Distributed Manufacturing - a geographically distributed network of facilities - has arisen as a promising option for supply-chain networks in which the transportation costs and infrastructure are the main bottlenecks (e.g., biomass (You and Wang, 2011), electric power, and shale gas) (Lara and Grossmann, 2016). However, despite the potential advantages of having distributed facilities, conventional large-scale centralized manufacturing can be more cost-effective due to economies of scale. Therefore, there is a need for a general optimization framework that can support the selection of centralized and distributed facilities taking into account the potential trade-offs (Lara and Grossmann, 2016).

We address the design and planning of manufacturing networks considering the selection and location in the continuous 2-dimensional space of centralized and/or distributed manufacturing facilities. The problem is formulated as a version of the continuous facility location-allocation problem with limited capacity.

1.1.2 Continuous Facility Location-Allocation Problem

The continuous facility location-allocation problem, also known as the Weber problem, has as its main objective to determine the locations in continuous 2-dimensional space for opening new facilities that are connected to supply and customer nodes, and the allocation of the material flows in this network, while minimizing the overall costs (Brimberg et al., 2008).

The Weber problem was named after Alfred Weber (Weber and Friedrich, 1929), whose work is considered to have established the foundations of modern location theories. In his first problem, he considers one facility to be located based on two suppliers and one customer, when these three points are not collinear (Weber and Friedrich, 1929; Khun and Kuenne, 1962). The basic model assumes Euclidean distances, but other distance functions such as Manhattan (or ℓ_1) norm and ℓ_p norm have also been used depending on the application and region where the transportation is considered. Advances in the continuous facility location-allocation with unlimited capacity can be found in the survey by Brimberg et al. (2008).

The capacitated version of the Weber problem, known as Capacitated Multi-facility Weber problem (CMWP), considers that the facilities to be installed have a maximum capacity. As shown by Sherali and Nordai (1988), this class of problems is NP-hard even if all the fixed points are located on a straight line. The general formulation for a CMWP is as follows.

$$\min \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} c_{k,j} \cdot f_{k,j} \cdot \left\{ (x_k - x_j)^2 + (y_k - y_j)^2 \right\}^{1/2} \tag{1.1a}$$

$$\text{s.t.} \quad \sum_j f_{k,j} = mc_k \qquad \forall\, k \in \mathcal{K} \tag{1.1b}$$

$$\sum_k f_{k,j} = d_j \qquad \forall\, j \in \mathcal{J} \tag{1.1c}$$

$$f_{k,j} \geq 0 \qquad \forall\, k \in \mathcal{K},\ j \in \mathcal{J} \tag{1.1d}$$

$$x_k, y_k \geq 0 \qquad \forall\, k \in \mathcal{K} \tag{1.1e}$$

where k is the index of the facilities to be located, j is the index of customers, (x_j, y_j) are the fixed coordinates of the customer j, mc_k is the capacity of facility k, d_j is the demand of customer j, and $c_{k,j}$ is the cost of unit flow per unit distance from facility k to customer j. The decision variables are: (x_k, y_k), which represent the coordinates of a new facility k; and $f_{k,j}$, which is the material flow between facility k to customer j.

Cooper (1972) was the first to attempt solving this type of location-allocation problem. He proposes exact and approximate solution methods based on the property that an optimal allocation occurs at the extreme point of the transportation polytope, while the optimal set of locations lies in the convex hull of the locations of the existing facilities. His exact formulation requires the explicit enumeration of the extreme points of the transportation polytope, limiting its application to small problems. Cooper's heuristic approach, known as the Alternating Transportation-Location (ATL) method, exploits the structure of the problem by alternating the solution of the transportation and allocation problems until convergence is achieved, but there is no guarantee of global optimality. Cooper (1975, 1976) further develops the ATL heuristic, extending it to CMWP with fixed charges.

Sherali and Shetty (1977) develop a cutting plane algorithm for the rectilinear distance location-allocation problem. Sherali and Tuncbilek (1992) propose a branch-and-bound algorithm for the squared-Euclidean distance location-allocation problem. Sherali et al. (2002) develop a branch-and-bound algorithm based on the partitioning of the allocation space that finitely converges to a global optimum within a tolerance. Chen et al. (2011) reformulate the problem as a sequence of nonlinear second order conic problems, and applied the semi-smooth Newton method to solve it. Akyuz et al. (2018) propose two branch-and-bound algorithms for solving exactly the multi-commodity CMWP: one based on partitioning the allocation space, and the other one considers partitioning of the location space. Besides the exact methods, there are several heuristics developed for this class of problem (Brimberg et al., 2008; Aras et al., 2007; Akyuz et al., 2012; Luis et al., 2015, 2016; Akyuz, 2017; Irawan et al., 2018).

1.1.3 Electric Power Infrastructure Planning

Changes in electricity demand, together with the wear-and-tear and retirement of old power plants, and the advances in the technology pool for electricity generation and storage, make it necessary to expand or adapt the electric power infrastructure. Generation Expansion Planning (GEP) models can be used to support the decision-making process in the power sector considering multiple energy sources (e.g., coal, natural gas, wind, solar), as well as to study the impact of new technology developments, resource cost trends, and policy shifts (e.g. carbon tax, and minimum renewable generation quota) (Sadeghi et al., 2017; Koltsaklis and Dagoumas, 2018; Babatunde et al., 2018; Gacitua et al., 2018).

Although transmission expansion is not considered in this work, it is important to be aware of its impact on long-term planning decisions, and thus we discuss it here. Traditionally, generation and transmission expansion are modeled separately: the generation is planned first and the transmission network is designed to meet this supply (e.g., Zhu and Chow (1997); Bahiense et al. (2001); Latorre et al. (2003); Alguacil et al. (2003); Bakirtzis et al. (2012)). Their simultaneous optimization is, however, a better way of capturing the trade-offs between investing in local generation or transmission from remote supplies (Krishnan et al., 2016). We recognize the potential benefits of co-optimizing transmission and generation expansion and that this co-optimized problem is important, especially in the context of high insertion of renewables (Krishnan et al., 2016). However, since these decisions are typically made independently of one another (Munoz et al., 2012), we chose to pursue the generation expansion side only, as has been done by many authors.

There is growing interest to use planning models to study scenarios with increasing penetration of solar and wind generation (MacDonald et al., 2016). Historically, since power systems were dominated by *dispatchable* thermal resources, planning models could ignore short-term operating constraints and have longer time periods without impacting much the quality of the results. Therefore, many GEP models (e.g. Regional energy deployment system (ReEDS) (Short et al., 2011), Integrated Planning Model (IPM) (EPA, 2015), US REGEN (EPRI, 2015)) approximate annual grid operations by using representative "time slices" to capture the average trends in load, wind and solar power output throughout the day and across seasons. Several widely-used multi-sector energy-economic models (e.g. NEMS (EIA, 2016)) and integrated assessment models (e.g. GCAM (Muratori et al., 2016)) also use this temporal representation to model power sector operation.

However, in a system deriving a large proportion of generation from intermittent resources, multiple analyses have indicated that such approximations may not be sufficient to represent the significant variability in renewable energy generation observed at hourly or sub-hourly time-scales (NERC, 2009; Albadi and El-Saadany, 2010; Lannoye et al., 2011; Nahmmacher et al., 2016b; Bistline et al., 2017; Deane et al., 2012), making it critical to consider hourly or sub-hourly operational decisions to assess the flexibility of the system (e.g., NERC (2009); Albadi and El-Saadany (2010); Lannoye et al. (2011)). Only then it is possible to rigorously assess the trade-offs between long-term investment decisions and short-term operating decisions. Accordingly, several papers have studied the impact of including operating constraints such as unit commitment, ramping limits and operating reserves in long term planning models (e.g., Shortt and O'Malley (2010); Ding and Somani (2010); Palmintier and Webster (2011); Pina et al. (2013); Mai et al. (2013); Palmintier and Webster (2014); Poncelet et al. (2014); Koltsaklis and Georgiadis (2015); Levin and Botterud (2015); Flores-Quiroz et al. (2016); de Sisternes et al. (2016); Heuberger et al. (2017); Oree et al. (2017); Mallapragada et al. (2018)).

Energy storage can play an important role in the future of renewable generation as they can smooth out the variability of wind and solar power output. A range of stationary, large-scale energy storage technologies are under development and, according to projections (Schmidt et al., 2017), they will significantly decrease their capital cost in the next few decades. Liu et al. (2017) consider a generic energy storage technology in their GEP model, which was represented as a multistage stochastic linear programming model, and applied it to a case-study in the Electric Reliability Council of Texas (ERCOT) region.

Power systems are subject to a variety of systematic uncertainties such as fuel prices, load demand, renewable generation, disruptive technologies, and future policies. However, because of the computational expense of combining uncertainty with a complete representation of the grid, and integrating detailed operating decisions with investment decisions over long planning horizons, most of the available commercial tools (Loulou et al., 2004, 2005; Short et al., 2011; EPRI, 2013; Diamant, 2017) and academic models (Ding and Somani, 2010; Short et al., 2011; Pina et al., 2013; Koltsaklis and Georgiadis, 2015; Flores-Quiroz et al., 2016; Heuberger et al., 2017; Lara et al., 2018a) are deterministic. The body of literature that addresses GEP optimization problem under uncertainty can be classified into two fundamental approaches for capturing uncertainty: Robust Optimization and Stochastic Programming.

The main idea behind Robust Optimization (RO) is to guarantee feasibility over a specified uncertainty set by modeling uncertain variables using bounds (Bertsimas and Sim, 2004). In general, this means that the computational burden of RO is much lower than that of stochastic programming. However, RO predicts more conservative results compared to the latter (Grossmann et al., 2016). Malcolm and Zenios (1994) were the first to propose a RO model for power systems capacity planning. Since then, other authors have explored GEP formulations in the context of RO (Mulvey and Ruszczyński, 1995; Chen et al., 2012; Li et al., 2014; Li, 2014). More recently, Adjustable Robust Optimization (ARO) has risen to prominence as an alternative to classic RO. ARO introduces recourse into the traditional RO formulation, allowing the model to respond to some uncertainty and generate less conservative solutions (Ben-Tal et al., 2004; Zhang et al., 2016). Some of the papers addressing GEP with ARO are Mejia-Giraldo (2013); Mejía-Giraldo and McCalley (2014); Moreira et al. (2017); Baringo and Baringo (2018).

Stochastic Programming (SP) is the most popular modeling framework for GEP problems under uncertainty. SP tends to be more appropriate for long-term production planning and strategic design decisions because it allows recourse decisions in the future to adapt to how the uncertainties are revealed (Grossmann et al., 2016). Thus, it is less conservative than RO. SP assumes uncertain data are random variables with known probability distributions and uses sampled values from this distribution to build a scenario tree and optimize over the expectation (Birge and Louveaux, 2011). As a downside, the solution is dependent on the accuracy of the assumed probability distributions of the uncertain parameters. However, there is mathematical theory and computational evidence that solutions obtained from SP are often stable with respect to changes in input probability distributions (Rachev and Römisch, 2002).

SP models can be formulated as two-stage and multistage problems. A typical two-stage stochastic GEP model considers as first-stage *here-and-now* decisions the investment decision over the entire planning horizon, which is made before the uncertainty realization. In this context, the second-stage *wait-and-see* decisions are the operational decisions, which are fully adaptive to the uncertainty realization. Some of the GEP literature that formulates the problem as two-stage stochastic programming includes: Dentcheva and Römisch (1998); Albornoz et al. (2004); Lopez et al. (2007); Jin et al. (2011); Wogrin et al. (2011); O'Neill et al. (2013); Gandulfo et al. (2014); Jin et al. (2014); Munoz and Watson (2015); Gil et al. (2015).

Multistage stochastic programming GEP models allow recourse between investment decisions in each stage, hence, they are also fully adaptive to the uncertainty realization. Park and Baldick (2016) propose a multistage stochastic mixed-integer program to solve GEP under load and wind availability uncertainty, and solve the model using a rolling-horizon. Zhan et al. (2017) propose a multistage stochastic programming model with endogenous uncertainty for GEPs with large amounts of wind power, and introduce a quasi-exact solution approach to reformulate the model as a mixed-integer linear programming model. Liu et al. (2018) propose a multistage linear stochastic GEP model that captures both large-scale uncertainties (e.g., investment, fuel-cost demand-growth rate uncertainty), and small-scale uncertainties (e.g., hourly demand and renewable generation uncertainty), and use progressive hedging algorithm to decompose the model by scenario and reduce computation times. Zou et al. (2018a) propose a partially adaptive stochastic mixed-integer optimization model in which the capacity expansion is fully adaptive to uncertainty up to a certain period and follows a two-stage approach thereafter, and propose an approximation algorithm to solve their model efficiently.

As mentioned before, even deterministic GEP models can pose significant computational challenges as the temporal and spatial scale resolution are increased (Lara et al., 2018a). The added complexity of handling uncertainty greatly intensifies this challenge, especially for multistage stochastic programming formulations as the scenario tree grows exponentially with the number of stages. Therefore, significant research has been devoted to the development of decomposition techniques to allow the solution of these problems in an efficient matter.

The most popular decomposition methods applied to multistage stochastic programming problems can be classified as scenario-based (e.g., Lagrangean Decomposition (Gupta and Grossmann, 2014), Dual Decomposition (Carøe and Schultz, 1999; Kim and Zavala, 2018), Augmented Lagrangean Mulvey and Ruszczyński (1995) and Progressive Hedging (Watson and Woodruff, 2011)), and stage-based decomposition (e.g., Nested Benders Decomposition (Birge, 1985), Stochastic Dual Dynamic Programming (SDDP) (Pereira and Pinto, 1991), Stochastic Dual Dynamic integer Programming (SDDiP)(Zou et al., 2018b)). Both categories have guaranteed finite convergence for Linear Programming (LP) formulations. However, for the case of Mixed-integer Linear Programming (MILP) formulations they can provide bounds to the optimal solution, but generally do not have guaranteed finite convergence.

Scenario-based decomposition often utilizes the framework of Lagrangean decomposition to decompose the problem into scenarios by dualizing the nonanticipativity constraints. Examples of GEP decompositions that fit within this category are: Jin et al. (2011); Munoz and Watson (2015); Liu et al. (2018). Stage-based decomposition decomposes the model by nodes in the scenario tree, and are usually based on Benders decomposition. Stage-based algorithms have smaller subproblems and are more suitable for models in which the solution of a single scenario is already computationally very demanding. SDDP has been widely used in the context of optimal scheduling of hydrothermal generating systems (Pereira and Pinto, 1985, 1991; Rebennack et al., 2012; Thome et al., 2013). Regarding GEP problems, both Nested Decomposition and SDDP have been used in combination with Benders decomposition for two-stage stochastic programming models in which the operational subproblems are challenging (Gorenstin et al., 1993; Rebennack, 2014). SDDiP (Zou et al., 2018b), which is an extension of SDDP to multistage integer programming models, is a promising technique to solve multistage stochastic integer programming models that has great potential for GEP models with operating details on an hourly basis.

1.2 Modeling frameworks and solution strategies

1.2.1 Generalized Disjunctive Programming (GDP)

Generalized disjunctive programming (GDP) is an extension of the disjunctive programming paradigm developed by Balas (1979). The GDP formulation involves algebraic equations, disjunctions and logic propositions in the formulation of a model, and makes the formulation process more intuitive and systematic, while retaining in the model the underlying logic structure of the problem (Grossmann and Trespalacios, 2013).

The general GDP formulation can be represented as follows:

$$\min \quad f(x) \tag{1.2a}$$

$$\text{s.t.} \quad g(x) \leq 0 \tag{1.2b}$$

$$\bigvee_{i \in D_k} \begin{bmatrix} Y_{ki} \\ r_{ki}(x) \leq 0 \end{bmatrix} \qquad k \in K \tag{1.2c}$$

$$\veebar_{i \in D_k} Y_{ki} \qquad k \in K \tag{1.2d}$$

$$\Omega(Y) = True \tag{1.2e}$$
$$x^{lo} \leq x \leq x^{up} \tag{1.2f}$$
$$x \in \mathbb{R}^n \tag{1.2g}$$
$$Y_{ki} \in \{True, False\} \qquad k \in K, i \in D_k \tag{1.2h}$$

The objective function (1.2a), which is a function of the continuous variables $x \in \mathbb{R}^n$, is subjected to a set of global constraints (1.2b) (i.e. constraints that must be satisfied regardless of the discrete decisions). The formulation involves disjunctions (1.2c), each of which contains disjunctive terms $i \in D_k$. The disjunctive terms in each disjunction are linked together by an "or" operator ($\vee$). A Boolean variable Y_{ki} and a set of constraints $r_{ki}(x) \leq 0$ are assigned to each disjunctive term. Exactly one disjunctive term in each disjunction must be enforced, as imposed by constraint (1.2d). A Boolean variable takes a value of $True$ ($Y_{ki} = True$) when a disjunctive term is active, and the corresponding constraints ($r_{ki}(x) \leq 0$) are enforced. When a term is not active ($Y_{ki} = False$), its corresponding constraints are ignored. The logic constraints (1.2e) represent the relations between the Boolean variables in propositional logic. Note that if there are equality constraints $g(x) = 0$, they can be represented by $g(x) \leq 0$ and $-g(x) \leq 0$.

GDP problems are typically reformulated as MILP/MINLP by using either the Big-M or Hull Reformulation (Wolsey and Nemhauser, 2014; Lee and Grossmann, 2000). The Big-M reformulation generates a smaller MINLP (i.e. less variables and constraints), while the Hull Reformulation provides a tighter formulation (Grossmann and Lee, 2003) since it represents the intersection of the convex hulls of each disjunction.

The Big-M reformulation is as follows:

$$\min \quad f(x) \tag{1.3a}$$
$$\text{s.t.} \quad g(x) \leq 0 \tag{1.3b}$$
$$r_{ki}(x) \leq M^{ki}(1 - y_{ki}) \qquad k \in K, i \in D_k \tag{1.3c}$$
$$\sum_{i \in D_k} y_{ki} = 1 \qquad k \in K \tag{1.3d}$$

$$Hy \geq h \tag{1.3e}$$
$$x^{lo} \leq x \leq x^{up} \tag{1.3f}$$
$$x \in \mathbb{R}^n \tag{1.3g}$$
$$y_{ki} \in \{0,1\} \qquad k \in K, i \in D_k \tag{1.3h}$$

In (1.3) the Boolean variables Y_{ki} are transformed into binary variables y_{ki}: $Y_{ki} = True$ is equivalent to $y_{ki} = 1$ and $Y_{ki} = False$ is equivalent to $y_{ki} = 0$. Constraint (1.3d) enforces that exactly one disjunctive term is selected per disjunction. The transformation of logic constraints $\Omega(Y) = True$ to integer linear constraints (1.3e) is easily obtained (Williams, 2013; Grossmann and Trespalacios, 2013). For an active term, the corresponding constraints $r_{ki}(x) \leq 0$ are enforced. For a term that is not active ($y_{ki} = 0$) and a large enough M^{ki}, the corresponding constraints $r_{ki}(x) \leq M^{ki}$ become redundant.

The Hull Reformulation (Grossmann and Trespalacios, 2013) is given as follows:

$$\min \quad f(x) \tag{1.4a}$$
$$\text{s.t.} \quad g(x) \leq 0 \tag{1.4b}$$
$$x = \sum_{i \in D_k} \nu^{ki} \qquad k \in K \tag{1.4c}$$
$$y_{ki} r_{ki}(\nu^{ki}/y_{ki}) \leq 0 \qquad k \in K, i \in D_k \tag{1.4d}$$
$$\sum_{i \in D_k} y_{ki} = 1 \qquad k \in K \tag{1.4e}$$
$$Hy \geq h \tag{1.4f}$$
$$x^{lo} y_{ki} \leq \nu^{ki} \leq x^{up} y_{ki} \qquad k \in K, i \in D_k \tag{1.4g}$$
$$x \in \mathbb{R}^n \tag{1.4h}$$
$$\nu^{ki} \in \mathbb{R}^n \qquad k \in K, i \in D_k \tag{1.4i}$$
$$y_{ki} \in \{0,1\} \qquad k \in K, i \in D_k \tag{1.4j}$$

In the Hull Reformulation, similarly to Big-M, the Boolean variables Y_{ki} are transformed into 0-1 variables y_{ki}, $\Omega(Y) = True$ is transformed into (1.4f), and (1.4e) enforces that only one disjunctive term is selected per disjunction. In the Hull Reformulation, the continuous variables x are disaggregated into variables ν^{ki}, for each disjunctive term $i \in D_k$ in each disjunction $k \in K$. The constraint (1.4g) enforces that when a term is active ($y_{ki} = 1$), the corresponding disaggregated variables lie within their bounds. When it is not selected, they take a value of zero. The constraint (1.4c) enforces that the original variables x have the same value as the disaggregated variables of the active terms. The functions in the constraints of a disjunctive term ($r_{ki}(x) \leq 0$) are represented by the perspective function $y_{ki}r_{ki}(\nu^{ki}/y_{ki})$ (Ceria and Soares, 1999) in constraint (1.4d). When a term is active ($y_{ki} = 1$) the constraint is enforced for the disaggregated variable ($r_{ki}(\nu^{ki}) \leq 0$). When it is not active ($y_{ki} = 0$), the constraint is trivially satisfied ($0 \leq 0$). When the constraints in the disjunction are linear ($A^{ki}x \leq a^{ki}$), the perspective function becomes $A^{ki}\nu^{ki} \leq a^{ki}y_{ki}$, which is a well-known representation in disjunctive programming ($Balas$, 1985). To avoid singularities in the perspective function, the following approximation can be used(Sawaya, 2006):

$$y_{ki}r_{ki}(\nu^{ki}/y_{ki}) \approx ((1-\epsilon)y_{ki} + \epsilon)r_{ki}\left(\frac{\nu^{ki}}{(1-\epsilon)y_{ki} + \epsilon}\right) - \epsilon r_{ki}(0)(1 - y_{ki}) \tag{1.5}$$

where ϵ is a small finite number (e.g. 10^-5). This approximation yields an exact value at $y_{ki} = 0$ and $y_{ki} = 1$ irrespective of the value of ϵ, and is convex if r_{ki} is convex.

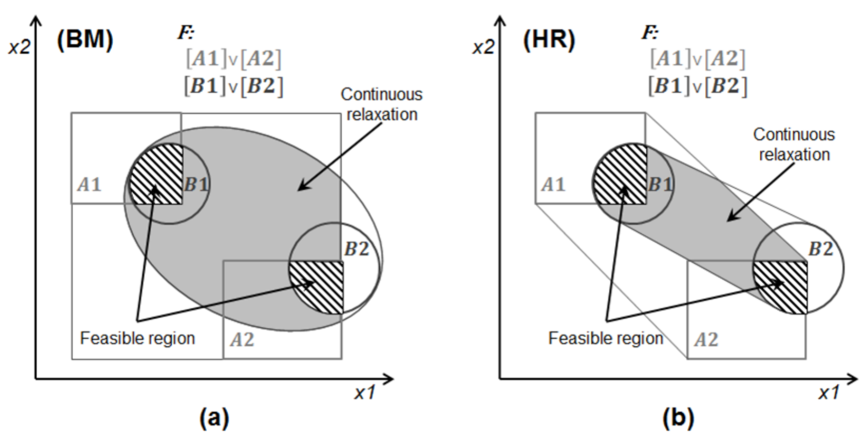

Figure 1.1: Illustration of Big-M (BM) and Hull Relaxation (HR) reformulations (Trespalacios, 2015)

Figure 1.1 illustrates the projection over $x1$ and $x2$ of the feasible region defined by two disjunctions for both reformulations (Trespalacios, 2015). The first disjunction represents the selection of rectangle $A1$ or rectangle $A2$, and the second one the selection of circle $B1$ or circle $B2$. The dashed region defines the feasible region, and the shaded area represents the continuous relaxation of the Big-M and Hull Relaxation. It is clear that the Hull Relaxation has a tighter relaxation than the Big-M.

1.2.2 Stochastic Programming

Optimization under uncertainty, also known as Stochastic Optimization, is an umbrella term that includes multiple communities, using different modeling styles, notation systems, and solution approaches (e.g., stochastic programming, robust optimization, decision trees, stochastic search, optimal control, Markov decision processes, approximate/adaptive/neuro-dynamic programming, reinforcement learning, model predictive control, simulation optimization) - the "jungle of stochastic optimization" (Powell, 2014). In this thesis we focus on the special case of Stochastic Programming (SP) and follow the same notation as this community.

The field of SP evolved from deterministic linear programming with the introduction of uncertain data as random variables with known probability distribution. In SP, the nature of the decisions is related to the knowledge of the realization of the uncertainty at a given *stage* in the decision-making process. Decisions are classified as *here-and-now*, when they are taken before the realization of the uncertainty, and *wait-and-see*, when taken after the values of the uncertain parameters (random variables) are revealed (i.e., recourse actions).

The classical two-stage stochastic linear programming formulation (Dantzig, 1955; Beale, 1955) is given as follows:

$$\min\ c^\intercal x + \mathbb{E}_\xi \left[\min q(\omega)^\intercal y(\omega)\right] \tag{1.6a}$$

$$\text{s.t.}\quad Ax = b \tag{1.6b}$$

$$T(\omega)x + Wy(\omega) = h(\omega) \qquad \forall \omega \in \Omega \tag{1.6c}$$

$$x \geq 0,\ y(\omega) \geq 0 \qquad \forall \omega \in \Omega \tag{1.6d}$$

where the first-stage decisions are represented by x. In the second stage, a number of random values $\omega \in \Omega$ may realize and, for a given realization ω, the second-stage problem data $q(\omega)$, $h(\omega)$ and $T(\omega)$ become known. Each component of q, T, and h is thus a possible random variable, and piecing together the stochastic components of the second-stage data, we obtain $\xi(\omega)$ (Birge and Louveaux, 2011).

The objective function (1.6a) contains a deterministic term $c^\intercal x$ and the expectation of the second-stage objective, $q(\omega)^\intercal y(\omega)$, taken over all realizations of the random event ω. Therefore, in this second-stage objective term, for each ω the value $y(\omega)$ is the solution of a linear program. To emphasize this fact, the SP formulation can be represented by its *deterministic equivalent* program. For a given realization ω, the second-stage value function is defined as follows:

$$Q(x, \xi(\omega)) = \left\{ \min_{y} q(\omega)^\intercal y : Wy = h(\omega) - T(\omega)x,\ y \geq 0 \right\} \tag{1.7}$$

Hence, the expected second-stage value function is

$$\mathcal{Q}(x) = \mathbb{E}_\xi Q(x, \xi(\omega)) \tag{1.8}$$

and the *deterministic equivalent* of problem (1.6) is defined as follows:

$$\min_{x \geq 0} \{c^\intercal x + \mathcal{Q}(x) : Ax = b\} \tag{1.9}$$

The general formulation of a two-stage integer program is very similar to the two-stage linear program (1.6). The only difference is that $x \geq 0,\ y(\omega) \geq 0$ is replaced by $x \in X,\ y(\omega) \in Y$ and X and/or Y contains some integrality or binary restrictions on x and/or y. However, if the integrality restrictions are present in the second stage, i.e., if the SP has mixed-integer recourse, the SP is much harder to solve because the expected recourse function $\mathcal{Q}(x)$ is, in general, lower semicontinuous, nonconvex and discontinuous (see Proposition 20 of Birge and Louveaux (2011)).

This two-stage decision-making process can be generalized to account for multiple stages.

Multistage stochastic programming models allow recourse between decisions in each stage. Hence, they are also fully adaptive to the uncertainty realization. The general multistage stochastic programming formulation is given as follows (Birge and Louveaux, 2011):

$$\min\ c_1x_1 + \mathbb{E}_{\xi_2}[\min c_2(\omega)x_2(\omega_2) + \ldots + \mathbb{E}_{\xi_H}[\min c_H(\omega)x_H(\omega_H)]\ldots] \tag{1.10a}$$

$$\text{s.t.}\quad W_1x_1 = h_1 \tag{1.10b}$$

$$T_1(\omega_2)x_1 + W_2x_2(\omega_2) = h_2(\omega) \tag{1.10c}$$

$$\ldots \tag{1.10d}$$

$$T_{H-1}(\omega_H)x_{H-1}(\omega_{H-1}) + W_Hx_H(\omega_H) = h_H(\omega) \tag{1.10e}$$

$$x_1 \in X_1; x_t(\omega_t) \in X_t, t = 2, ..., H; \tag{1.10f}$$

The *deterministic equivalent* of the general multi-stage stochastic programming is then defined as follows (Birge and Louveaux, 2011):

$$\min_{x\in X_1} \{c_1x_1 + \mathcal{Q}_2(x_1) : W_1x_1 = h_1\} \tag{1.11}$$

where the expected value function for stage $t+1$ is given by:

$$\mathcal{Q}_{t+1}(x_t) = \mathbb{E}_{\xi_{t+1}}[Q_{t+1}(x_t, \xi_{t+1}(\omega))] \tag{1.12}$$

for all t to obtain the recursion for $t = 2, ..., H-1$,

$$Q_t(x_{t-1}, \xi_t(\omega)) = \left\{ \min_{x(\omega)\in X_t}\ c_t(\omega)x_t(\omega) + \mathcal{Q}_{t+1}(x_t) : W_tx_t(\omega) = h_t(\omega) - T_{t-1}(\omega)x_{t-1} \right\} \tag{1.13}$$

For terminal conditions $t = H$, we have:

$$Q_H(x_{H-1}, \xi_H(\omega)) = \left\{ \min_{x(\omega)\in X_H}\ c_H(\omega)x_H(\omega) : W_Hx_H(\omega) = h_H(\omega) - T_{H-1}(\omega)x_{H-1} \right\} \tag{1.14}$$

When we have a finite number of possible realizations of the future outcomes, the set of possible future sequences of outcomes are called *scenarios*. The description of scenarios is often made on a scenario trees such as that in Figure 1.2. These problems become extremely large as the number of stages increases, even if only a few realizations are allowed in each stage, as the scenario tree grows exponentially with the number of stages.

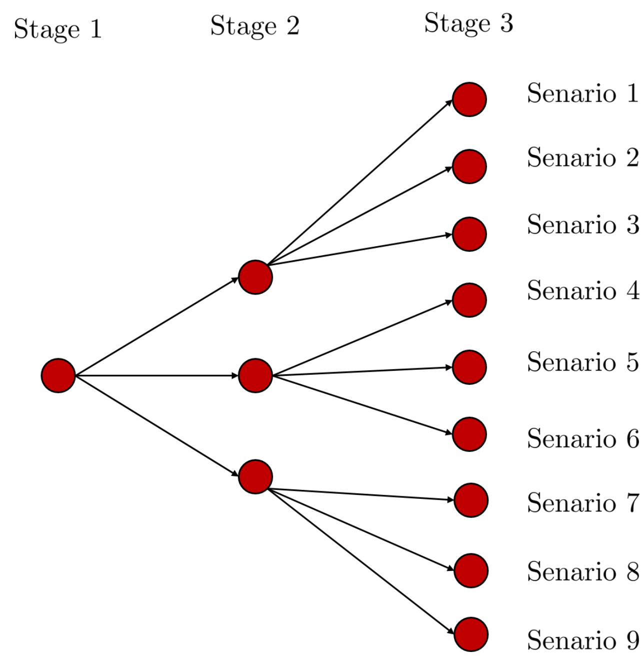

Figure 1.2: Illustration of a scenario tree with 3 stages and 3 realizations per stage

1.2.3 Decomposition methods for mathematical optimization

To take advantage of special structures in the formulation of an optimization problem, it may be desirable to use a decomposition algorithm. Decomposition methods can be:

- *ad hoc*, i.e. created for a particular problem to take advantage of its specific properties (which is the case of the Bilevel Decomposition and Accelerated Bilevel Decomposition proposed in Part I of this thesis).
- general, i.e. to target a class of problems with similar structures (which is the case of the Nested Decomposition algorithm and Stochastic Dual Dynamic Programming introduced in Part II of this thesis).

Regarding the general decomposition algorithms, they can be classified according to their structure into *complicating constraints* methods and *complicating variable* methods (Conejo et al., 2006), as shown in the Figure 1.3. Note that complicating constraints involve variables from different blocks, and complicating variables link constraints pertaining to different blocks.

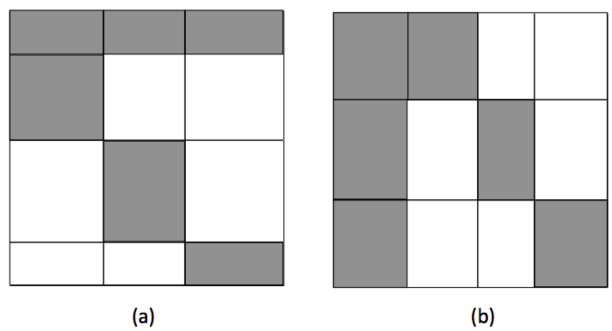

Figure 1.3: Schematic of the set of constraints of a decomposable optimization problem with complicating (a) constraints and (b) variables (Calfa, 2015).

If the constraints can be partitioned into a set of easily decomposable constraints and *complicating constraints*, a Lagrangean or a Dantzig-Wolfe algorithms may be appropriate. On the other hand, if the variables can be partitioned into easily decomposable variables and *complicating variables*, Benders Decomposition may be the most suitable approach (Conforti et al., 2014).

The remainder of this section gives an overview of Lagrangean Decomposition, Benders Decomposition, and Nested Benders Decomposition, which served as basis for the the general decomposition algorithms presented by this thesis.

Lagrangean Decomposition

Consider the following MILP formulation:

$$\begin{aligned}
\Phi = \quad & \max \quad cx && (1.15a)\\
& \text{s.t.} \quad Ax \le b \qquad \leftarrow \lambda \ge 0 && (1.15b)\\
& \qquad\quad Dx \le e && (1.15c)\\
& \qquad\quad x \in \mathbb{Z}_+^n && (1.15d)
\end{aligned}$$

where (1.15b) are complicating constraints, and $\lambda \geq 0$ are the Lagrange multipliers of the complicating constraints (1.15b). The Lagrangian relaxation (Geoffrion, 1974) of the MILP problem (1.15) is given by:

$$\Phi^{\mathrm{LR}}(\lambda) = \quad \max \quad cx - \lambda(b - Ax) \tag{1.16a}$$

$$\text{s.t.} \quad Dx \leq e \tag{1.16b}$$

$$x \in \mathbb{Z}_+^n \tag{1.16c}$$

and yields upper bounds to the original MILP problem (1.15), $\Phi^{\mathrm{LR}}(\lambda) \geq \Phi$.

The Lagrangean Dual is given by the following *min-max* problem:

$$\Phi^{\mathrm{Dual}} = \left\{ \min_{\lambda} \Phi^{\mathrm{LR}}(\lambda) : \lambda \geq 0 \right\} \tag{1.17a}$$

$$= \left\{ \min_{\lambda} \max_{x} x - \lambda(b - Ax) : Dx \leq e; \lambda \geq 0; x \in \mathbb{Z}_+^n \right\} \tag{1.17b}$$

where the minimization of the Lagrange multipliers is often obtained by using the subgradient method (Fisher, 2004).

Lagrangean Decomposition is a special case of Lagrangean Relaxation in which we define a different set of variables for each set of constraints (i.e., easily decomposable and complicating constraints), and (1.18d) becomes the new complicating constraints.

$$\Phi = \quad \max \quad cx \tag{1.18a}$$

$$\text{s.t.} \quad Ax \leq b \tag{1.18b}$$

$$Dy \leq e \tag{1.18c}$$

$$x = y \qquad \leftarrow \lambda \in \mathbb{R}^n \tag{1.18d}$$

$$x, y \in \mathbb{Z}_+^n \tag{1.18e}$$

where $\lambda \in \mathbb{R}^n$ are the Lagrange multipliers of constraint (1.18d).

The Lagrangean relaxation of (1.18) trivially decomposes into subproblems:

$$
\begin{aligned}
\Phi_1^{\mathrm{LR}}(\lambda) = \quad & \max \quad (c-\lambda)x & \qquad \Phi_2^{\mathrm{LR}}(\lambda) = \quad & \max \quad \lambda y && (1.19a)\\
& \text{s.t.} \quad Ax \le b & & \text{s.t.} \quad Dy \le e && (1.19b)\\
& \qquad x \in \mathbb{Z}_+^n & & \qquad y \in \mathbb{Z}_+^n && (1.19c)
\end{aligned}
$$

and the Lagrangean Dual is given by the following minimization:

$$
\Phi^{\mathrm{LD}} = \left\{ \min_{\lambda} \left(\Phi_1^{\mathrm{LR}}(\lambda) + \Phi_2^{\mathrm{LR}}(\lambda) \right) : \lambda \in \mathbb{R}^n \right\} \tag{1.20}
$$

The bound predicted by the Lagrangean decomposition is at least as tight as the one provided by Lagrangean Relaxation (Guignard and Kim, 1987), $\Phi^{\mathrm{LR}}(\lambda) \ge \Phi^{\mathrm{Dual}} \ge \Phi^{\mathrm{LD}} \ge \Phi$.

Benders Decomposition

The Benders Decomposition was proposed by Benders (1962), with the main objective of tackling problems with complicating variables, which, when temporarily fixed, yield a problem significantly easier to handle.

Consider the following MILP formulation:

$$
\begin{aligned}
\Phi = \quad & \max \quad f^{\mathsf{T}} y + c^{\mathsf{T}} x && (1.21a)\\
& \text{s.t.} \quad Ax = b && (1.21b)\\
& \qquad By + Dx = d && (1.21c)\\
& \qquad x \ge 0 && (1.21d)\\
& \qquad y \in \mathbb{Z}_+^n && (1.21e)
\end{aligned}
$$

where y are the complicating variables. Model (1.21) can be re-expressed as:

$$
\Phi = \min_{\hat{y} \in Y} f^{\mathsf{T}} \hat{y} + \min_{x \ge 0} \{ c^{\mathsf{T}} x : Dx = d - B\hat{y} \} \tag{1.22}
$$

where $\hat{y}$ is a given value for the complicating variables, which belongs to the set $Y = \{y | Ay = b, y \in \mathbb{Z}_+^n\}$.

Based on duality theory, the primal and dual formulations can be interchanged (Rahmaniani et al., 2016) to extract the following equivalent formulation:

$$\Phi = \min_{\hat{y} \in Y} f^\intercal \hat{y} + \max_{\lambda \in \mathbb{R}^n} \{\lambda^\intercal (d - B\hat{y}) : \lambda^\intercal D \leq c\} \tag{1.23}$$

where $\lambda \in \mathbb{R}^n$ are the Lagrange multipliers of constraints $Dx = d - B\hat{y}$.

The feasible space of the inner maximization, i.e., $F = \{\lambda | \lambda^\intercal D \leq c\}$, is independent of the choice of $\hat{y}$. Thus, if F is not empty, the inner problem can be either unbounded or feasible for any arbitrary choice of $\hat{y}$. Given the set of extreme rays Q of F, there is a direction of unboundedness $r_q, q \in Q$ for which $r_q^\intercal (d - B\hat{y}) > 0$; this must be avoided because it indicates the infeasibility of the $\hat{y}$ solution. If we add all the cuts of the form $r_q^\intercal (d - B\hat{y}) \leq 0 \; \forall q \in Q$ to the outer minimization problem, the value of the inner problem will be one of its extreme points (Rahmaniani et al., 2016). Consequently, problem (1.23) can be reformulated as the *Benders Master Problem*, making use of the continuous variable $\alpha \in \mathbb{R}^1$ to linearize the inner maximization problem:

$$\Phi^{\mathrm{UB}} = \min_{y,\alpha} f^\intercal y + \alpha \tag{1.24a}$$

$$\text{s.t.} \quad Ay = b \tag{1.24b}$$

$$\alpha \geq \lambda_e^\intercal (d - By) \qquad \forall e \in E \tag{1.24c}$$

$$0 \leq r_q^\intercal (d - By) \qquad \forall q \in Q \tag{1.24d}$$

$$y \in \mathbb{Z}_+^n \tag{1.24e}$$

where (1.24c) are the optimality cuts and (1.24d) are the feasibility cuts.

The *Benders Subproblem* is the dual formulation of problem (1.21) for a trial value of $\hat{y}$:

$$\Phi^{\mathrm{LB}} = \max_{\lambda \in \mathbb{R}^n} \{\lambda^\intercal (d - B\hat{y}) : \lambda^\intercal D \leq c\} \tag{1.25}$$

The Benders decomposition algorithm consists of solving the Master Problem (1.24) and the Subproblem (1.25) iteratively (starting with the Master Problem) until an optimal solution is found $\Phi^{\text{LB}} = \Phi^{\text{UB}}$. It is important to highlight that the Subproblem needs to be convex in order to generate valid cuts.

Nested Benders Decomposition

The nested Benders decomposition method is based on the idea of applying the Benders Decomposition method to a problem more than once, in a nested fashion. It was first proposed by Birge (1985) and it is particularly appropriate for multi-period/multi-stage problems in which each pair of adjacent periods/stages can be considered separately as a subproblem.

Consider the following multi-period LP problem where $t \in \{1, ..., T\}$ is the set of periods:

$$
\begin{aligned}
\Phi = \quad & \min \quad \sum_{t=1}^{T} c_t^{\intercal} x_t && && (1.26a)\\
\text{s.t.} \quad & A_t x_t = b_t && \forall t \in \{1, ..., T\} && (1.26b)\\
& B_t x_t + D_t x_{t-1} = d_t && \forall t \in \{2, ..., T\} && (1.26c)\\
& x_t \geq 0 && \forall t \in \{1, ..., T\} && (1.26d)
\end{aligned}
$$

This problem would be fully separable if it was not for constraint (1.26c) that links different periods (complicating constraint). Thus, following a similar procedure as the Lagrangean Decomposition, we define a new set of variables $z \geq 0$, and enforce if to have the same value as the trial value for x_t obtained in the previous period $z_t = \hat{x}_{t-1}$.

$$
\begin{aligned}
\Phi = \quad & \min \quad \sum_{t=1}^{T} c_t^{\intercal} x_t && && && (1.27a)\\
\text{s.t.} \quad & A_t x_t = b_t && && \forall t \in \{1, ..., T\} && (1.27b)\\
& B_t x_t + D_t z_t = d_t && && \forall t \in \{2, ..., T\} && (1.27c)\\
& z_t = \hat{x}_{t-1} && \leftarrow \lambda_t \in \mathbb{R}^n && \forall t \in \{2, ..., T\} && (1.27d)\\
& x_t, z_t \geq 0 && && \forall t \in \{1, ..., T\} && (1.27e)
\end{aligned}
$$

where $\lambda_t \in \mathbb{R}^n$ are the Lagrange multipliers of constraint (1.27d).

The problem is now trivially separable and the subproblem is given by (1.28) if $t = 1$, by (1.29) if $t = \{2, ..., T-1\}$, and by (1.30) if $t = T$:

$$\Phi_1 = \min \quad c_1^\intercal x_1 + \alpha_1 \tag{1.28a}$$
$$\text{s.t.} \quad A_1 x_1 = b_1 \tag{1.28b}$$
$$\alpha_1 \geq \hat{\Phi}_2 + \lambda_2(\hat{x}_1 - x_1) \tag{1.28c}$$
$$x_1 \geq 0 \tag{1.28d}$$

$$\left.\begin{aligned} \Phi_t = \min \quad & c_t^\intercal x_t + \alpha_t \\ \text{s.t.} \quad & A_t x_t = b_t \\ & B_t x_t + D_t z_t = d_t \\ & z_t = \hat{x}_{t-1} \qquad \leftarrow \lambda_t \in \mathbb{R}^n \\ & \alpha_t \geq \hat{\Phi}_{t+1} + \lambda_{t+1}(\hat{x}_t - x_t) \\ & x_t, z_t \geq 0 \end{aligned}\right\} \forall t \in \{2, ..., T-1\} \tag{1.29}$$

$$\Phi_T = \min \quad c_T^\intercal x_T \tag{1.30a}$$
$$\text{s.t.} \quad A_T x_T = b_T \tag{1.30b}$$
$$B_T x_T + D_T z_T = d_T \tag{1.30c}$$
$$z_T = \hat{x}_{T-1} \qquad \leftarrow \lambda_T \in \mathbb{R}^n \tag{1.30d}$$
$$x_T, z_T \geq 0 \tag{1.30e}$$

where $\hat{\Phi}_{t+1}$ is the optimal value obtained by the solution of the Subproblem of the period immediately after $t+1$, $\hat{x}_t$ are the trial values for x_t to be fixed when solving the period immediately after $t+1$, $\lambda_t \in \mathbb{R}^n$ is the set of Lagrange multipliers of the complicating linking constraint (1.27d), and α_t is the continuous variable used in the optimality cut similarly to the classic Benders Decomposition.

The nested Benders decomposition algorithm consists of solving the Subproblems in a forward (i.e. $t = \{1, ..., T\}$) and backward (i.e. $t = \{T, ..., 1\}$) fashion. The Forward Pass generates trial solutions for x_t to be fixed in the period after, and yields a feasible upper bound to the original LP problem (1.26). The Backward Pass generates the coefficients for the optimality cuts to be added in the subproblems of the period before, and the solution of period $t = 1$ yields a lower bound to the original LP problem (1.26). The cuts generated in the Backward Pass are kept in the formulation of the next Forward and Backward passes, which are solved iteratively until an optimal solution is found.

1.3 Overview of the thesis

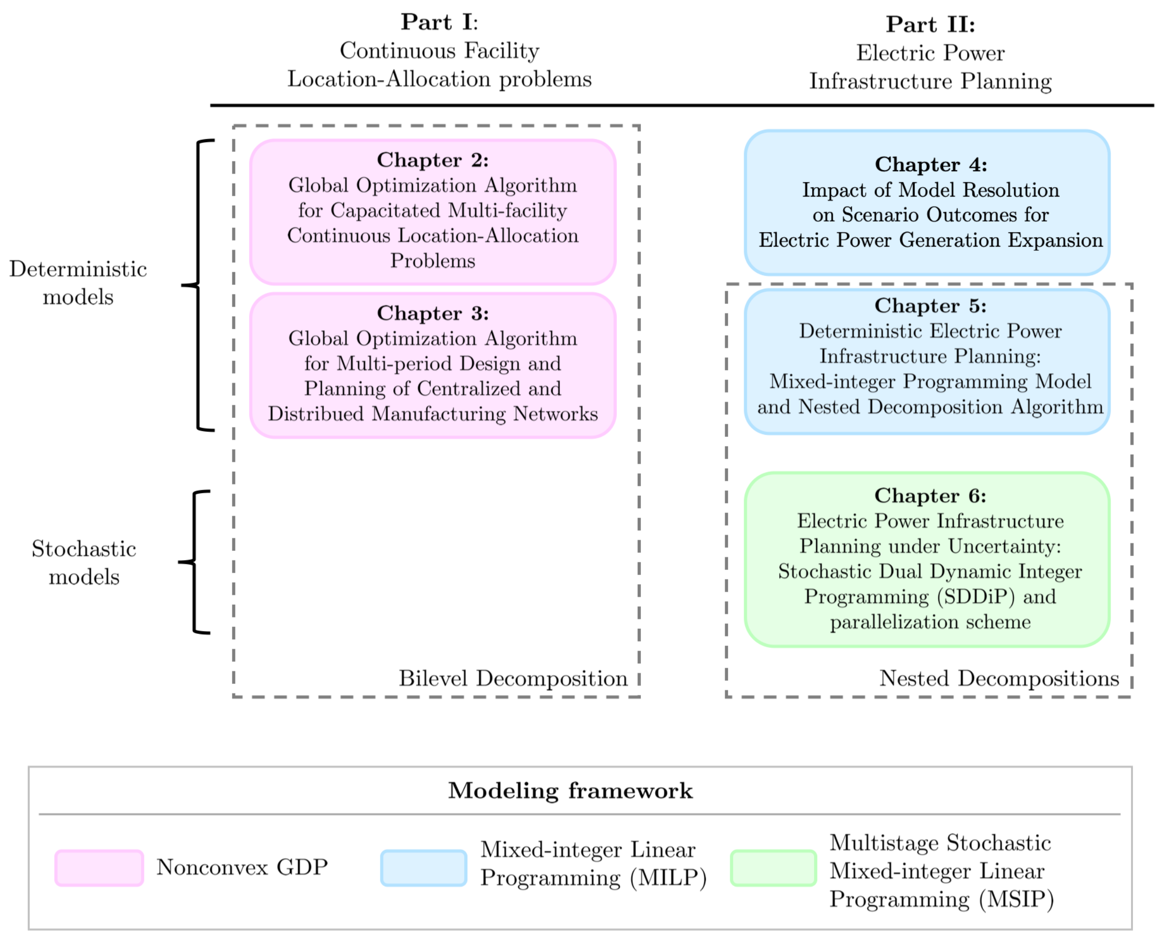

Figure 1.4: Overview of the thesis.

Figure 1.4 shows a graphic overview of the topics in Chapters 2-6. The following subsections provide an abstract of all the chapters.

Chapter 2

In chapter 2, we propose a nonconvex Generalized Disjunctive Programming (GDP) model to optimize the 2-dimensional continuous location-allocation of the potential facilities based on the maximum capacity and the given coordinates of the suppliers and customers. The model belongs to the class of Capacitated Multi-facility Weber Problem. We propose a bilevel decomposition algorithm that iteratively solves a discretized MILP version of the model, and its nonconvex NLP for a fixed selection of discrete variables. Based on the bounding properties of the subproblems, ϵ-convergence is proved for this algorithm. We apply the proposed method to random instances varying from 2 suppliers and 2 customers to 40 suppliers and 40 customers, from one type of facility to 3 different types, and from 2 to 32 potential facilities. The results show that the algorithm is more effective at finding global optimal solutions than general purpose global optimization solvers tested.

Chapter 3

In chapter 3, we address the design and planning of manufacturing networks considering the option of centralized and/or distributed facilities, taking into account the potential trade-offs. The problem is formulated as a version of the continuous facility location-allocation problem with limited capacity proposed in chapter 2, which involves the selection of which facilities to build in each time period, and their location in the continuous 2-dimensional space in order to meet demand and maximize profits. The model is a multi-period nonconvex Generalized Disjunctive Programming (GDP), reformulated as a multi-period nonconvex Mixed-Integer Nonlinear Programming (MINLP). We propose an accelerated version of the Bilevel Decomposition proposed in chapter 2 with additional steps to reduce the feasible space and help the accelerate the solution of the Master Problem (MILP). We benchmark the performance of the Accelerated Bilevel Decomposition algorithm against the original Bilevel Decomposition and the commercial global solvers available, and show that our proposed algorithm outperforms the other options in all instances tested. Additionally, we illustrate the applicability of the proposed model and solution framework with a case study for a biomass supply chain.

Chapter 4

In chapter 4, we investigate the impact of embedding additional operational and temporal detail in a Generation Expansion Planning (GEP) framework on the resulting projections for generation capacity additions and their utilization. This preliminary analysis provides motivation for the modeling and algorithmic efforts in the remaining chapters. Our approach is based on systematically comparing the outputs from a chronological hourly GEP with outputs from a commonly used time-slice GEP, using seasonally-averaged time blocks. The GEP models mainly differ in their representation of operational flexibility of thermal generators as well as the temporal resolution of load and renewable generation, and are both less complex than the models proposed in the remaining chapters. Studying the Texas grid over a range of hypothetical renewable energy penetration scenarios, we find that the time-slice approach tends to overestimate solar capacity and underestimate wind and natural gas capacity relative to the chronological approach. We also test capacity projections of both GEPs through an hourly grid operations model to explore operational metrics, such as the ability to meet demand subject to intra and inter-annual variations in load and renewable generation. This experiment reveals that the projections made by the chronological GEP consistently lead to lower unmet demand compared to the time-slice capacity projections. These findings imply the need for sufficient temporal resolution and chronology or validated parameterizations that yield similar behavior to be included in power sector GEPs and multi-sector energy-economic models using a time slice representation.

Chapter 5

Chapter 5 addresses the long-term planning of electric power infrastructure considering fossil fuels (coal, natural gas and nuclear) and high renewable penetration (wind and solar).

To capture the intermittency of the renewable sources, we propose a deterministic multi-scale Mixed-Integer Linear Programming (MILP) formulation that simultaneously considers annual generation investment decisions and hourly operational decisions. We adopt judicious approximations and aggregations to improve its tractability. Moreover, to overcome the computational challenges of treating hourly operational decisions within a monolithic multi-year planning horizon, we propose a decomposition algorithm based on Nested Benders Decomposition for multi-period MILP problems to allow the solution of larger instances. Our decomposition adapts previous nested Benders methods by handling integer and continuous state variables, although at the expense of losing its finite convergence property due to potential duality gap. We apply the proposed modeling framework to a case study in the Electric Reliability Council of Texas (ERCOT) region, and demonstrate massive computational savings from our decomposition.

Chapter 6

In chapter 6, we address the long-term planning of electric power infrastructure under uncertainty. We propose a Multistage Stochastic Mixed-integer Programming formulation that optimizes the generation expansion to meet the projected electricity demand over multiple years while considering detailed operational constraints, intermittency of renewable generation, power flow between regions, storage options, and multiscale representation of uncertainty (strategic and operational). To be able to solve this large-scale model, which grows exponentially with the number of stages in the scenario tree, we decompose the problem using Stochastic Dual Dynamic Integer Programming (SDDiP). The SDDiP algorithm is computationally expensive but we take advantage of parallel processing to solve it more efficiently. The proposed formulation and algorithm are applied to a case study in the region managed by the Electric Reliability Council of Texas (ERCOT) for scenario trees considering natural gas price and carbon tax uncertainty for a reference case and a hypothetical case without nuclear power. We show that the parallelized SDDiP algorithm allows the solution of instances with quadrillions of variables and constraints in reasonable amounts of time.

Chapter 7

Chapter 7 provides a critical review of the work in this thesis, along with a summary of its contributions and suggestions for future work.

Part I

Continuous Facility Location-Allocation Design and Planning

Chapter 2

Global Optimization Algorithm for Capacitated Multi-facility Continuous Location-Allocation Problems

In this chapter, we address the design of networks that involves the selection and location of facilities in the continuous 2-dimensional space. The problem is formulated as a continuous facility location-allocation problem with limited capacity, also known as the Capacitated Multi-facility Weber Problem (CMWP) with fixed costs. The objective of this type of problem is to determine locations in continuous 2-dimensional space for opening new facilities that are connected to supply and customer nodes, taking into account limited capacities and transportation costs (Brimberg et al., 2008).

We propose an extension of the Capacitated Multi-facility Weber Problem (CMWP) that considers fixed costs for opening new facilities, fixed transportation costs, and two sets of fixed-location points: suppliers i and customers j, as represented in Figure 2.1. The latter goes back to the original Weber problem, in which the location of the facility had to be determined in relation to 2 suppliers and 1 customer points. The model is a nonconvex Mixed-Integer Nonlinear Programming (MINLP), in which the nonconvexity comes from the variable multiplication in the transportation cost. This is, to the best of our knowledge, an original problem not reported before that has high practical applicability (Lara and Grossmann, 2016).

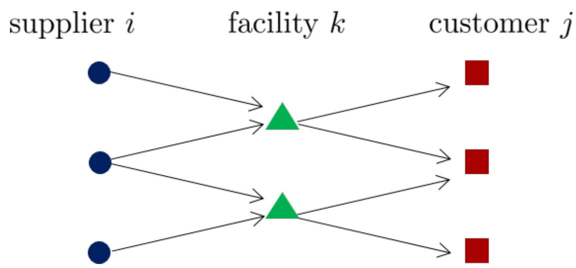

Figure 2.1: Representation of the nodes in the network

The remainder of the chapter is organized as follows. We begin by presenting in Section 2.1 the problem statement, its General Disjunctive Programming (GDP) formulation, and its reformulation as a nonconvex MINLP. In Section 2.2 we propose a global optimization algorithm based on the partitioning of the space, which is guaranteed to have ϵ-convergence. In Section 2.3 we introduce a small test problem. The performance of the algorithm is assessed in Section 2.4 by solving randomly generated instances with the proposed algorithm and comparing the solution, optimality gap and computational time with general purpose global optimization solvers.

2.1 Problem statement

Given is a set of suppliers $i \in \mathcal{I}$, with their respective fixed locations (x_i, y_i), availability a_i, and cost of material supply cs_i. Given is also a set of customers $j \in \mathcal{J}$, with their respective fixed locations (x_j, y_j), and demands d_j. Given are the fixed and variable costs (ff_k and vf_k, respectively) of potential facilities $k \in \mathcal{K}$ with N different types, which are divided in subsets $\mathcal{K}_n \; \forall \; n = \{1, ..., N\}$ such that $\bigcup_n \mathcal{K}_n = \mathcal{K}$. The corresponding maximum capacity, mc_k, and conversion to product flows, cv_k, of these potential facilities are also known. Given are also the transportation costs between suppliers and facilities, and facilities and customers ($ft_{i,k}, ft_{k,j}$: fixed costs; $vt_{i,k}, vt_{k,j}$: variable costs). The problem is to find the optimal network of facilities (number, types, location, and corresponding flows) that minimizes the total cost.

The variables in the problem are the coordinates of potential facilities, (x_k, y_k), the distances between supplier and facility, $D_{i,k}$, and between facility and customer, $D_{k,j}$, the flows between supplier and facility, $f_{i,k}$, and between facility and customer, $f_{k,j}$, and the amount produced by each facility, f_k. There are also Boolean variables W_k ($True$ if facility is built; $False$ otherwise); $Z_{i,k}$ ($True$ if material supply is transported between supplier and facility; $False$ otherwise); and $Z_{k,j}$ ($True$ if product is transported between facility and customer; $False$ otherwise). The GDP (Trespalacios and Grossmann, 2014) formulation is given by Equations (2.1a)-(2.1m).

$$\min \Phi = \sum_k Cost_k + \sum_i \sum_k Cost_{i,k} + \sum_k \sum_j Cost_{k,j} \tag{2.1a}$$

$$\text{s.t.} \begin{bmatrix} W_k \\ Cost_k = ff_k + vf_k \cdot f_k \\ 0 \leq f_k \leq mc_k \\ 0 \leq x_k \leq x_k^{\mathrm{U}} \\ 0 \leq y_k \leq y_k^{\mathrm{U}} \end{bmatrix} \vee \begin{bmatrix} \neg W_k \\ Cost_k = 0 \\ f_k = 0 \\ x_k = 0 \\ y_k = 0 \end{bmatrix} \qquad \forall\, k \in \mathcal{K} \tag{2.1b}$$

$$\begin{bmatrix} Z_{i,k} \\ Cost_{i,k} = cs_i \cdot f_{i,k} + ft_{i,k} + vt_{i,k} \cdot f_{i,k} \cdot D_{i,k} \\ 0 \leq f_{i,k} \leq f_{i,k}^{\mathrm{U}} \\ D_{i,k}^{\mathrm{L}} \leq D_{i,k} \leq D_{i,k}^{\mathrm{U}} \end{bmatrix} \vee \begin{bmatrix} \neg Z_{i,k} \\ Cost_{i,k} = 0 \\ f_{i,k} = 0 \end{bmatrix} \qquad \forall\, i \in \mathcal{I}, k \in \mathcal{K} \tag{2.1c}$$

$$\begin{bmatrix} Z_{k,j} \\ Cost_{k,j} = ft_{k,j} + vt_{k,j} \cdot f_{k,j} \cdot D_{k,j} \\ 0 \leq f_{k,j} \leq f_{k,j}^{\mathrm{U}} \\ D_{k,j}^{\mathrm{L}} \leq D_{k,j} \leq D_{k,j}^{\mathrm{U}} \end{bmatrix} \vee \begin{bmatrix} \neg Z_{k,j} \\ Cost_{k,j} = 0 \\ f_{k,j} = 0 \end{bmatrix} \qquad \forall\, k \in \mathcal{K}, j \in \mathcal{J} \tag{2.1d}$$

$$D_{i,k} \geq \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2} \qquad \forall\, i \in \mathcal{I}, k \in \mathcal{K} \tag{2.1e}$$

$$D_{k,j} \geq \sqrt{(x_j - x_k)^2 + (y_j - y_k)^2} \qquad \forall\, k \in \mathcal{K}, j \in \mathcal{J} \tag{2.1f}$$

$$W_k \iff \bigvee_i Z_{i,k} \qquad \forall\, k \in \mathcal{K} \tag{2.1g}$$

$$W_k \iff \bigvee_j Z_{k,j} \qquad \forall\, k \in \mathcal{K} \tag{2.1h}$$

$$\sum_k f_{i,k} \leq a_i \qquad \forall\, i \in \mathcal{I} \tag{2.1i}$$

$$\sum_i f_{i,k} \cdot cv_k = f_k \qquad \forall\, k \in \mathcal{K} \tag{2.1j}$$

$$f_k = \sum_j f_{k,j} \qquad \forall\, k \in \mathcal{K} \tag{2.1k}$$

$$\sum_k f_{k,j} = d_j \qquad \forall\, j \in \mathcal{J} \tag{2.1l}$$

$$W_k, Z_{i,k}, Z_{k,j} \in \{True, False\} \qquad \forall\, i \in \mathcal{I}, k \in \mathcal{K}, j \in \mathcal{J} \tag{2.1m}$$

The objective function (2.1a) includes costs for the facilities and for the transportation from the suppliers and to the customers. Disjunction (2.1b) determines the selection of facilities (W_k), while the disjunctions (2.1c) and (2.1d) determine the transportation links $\{i,\ k\}$ and $\{k,\ j\}$ with the corresponding Boolean variables ($Z_{i,k},\ Z_{k,j}$). Constraints (2.1e) and (2.1f) represent the Euclidean distances between suppliers and facilities, and facilities and customers, while the logic relations in (2.1g) and (2.1h) establish the existence of links depending on the choice of the facilities and vice-versa. Finally, constraints (2.1i)-(2.1l) define the mass balances as well as the availabilities and demands.

The model (2.1) is a nonconvex GDP due to the bilinear terms ($f \cdot D$) in the transportation cost, as can be seen in disjunctions (2.1c)-(2.1d). Equations (2.1e)-(2.1f) are nonlinear convex constraints since they correspond to Euclidean norms (Al-Loughani, 1997). The presence of nonconvexities was the main motivation for representing the problem as a GDP. By having the bilinear terms as part of the disjunctions, the transportation costs are calculated only for the selected connections within an iterative procedure.

The GDP can be transformed into an MINLP using the hull reformulation, which yields the tightest relaxation for each disjunction (Trespalacios and Grossmann, 2014). Since the disaggregated variables can be reformulated back to the original variables, the resulting MINLP is given by Equations (2.2a)-(2.2v).

$$\min\ \Phi = \sum_k Cost_k + \sum_i \sum_k Cost_{i,k} + \sum_k \sum_j Cost_{k,j} \tag{2.2a}$$

$$
\begin{aligned}
\text{s.t. } & Cost_k = ff_k \cdot w_k + vf_k \cdot f_k && \forall\ k \in \mathcal{K} && (2.2\text{b}) \\
& Cost_{i,k} = cs_i \cdot f_{i,k} + ft_{i,k} \cdot z_{i,k} + vt_{i,k} \cdot f_{i,k} \cdot D_{i,k} && \forall\ i \in \mathcal{I}, k \in \mathcal{K} && (2.2\text{c}) \\
& Cost_{k,j} = ft_{k,j} \cdot z_{k,j} + vt_{k,j} \cdot f_{k,j} \cdot D_{k,j} && \forall\ k \in \mathcal{K}, j \in \mathcal{J} && (2.2\text{d}) \\
& D_{i,k} \geq \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2} && \forall\ i \in \mathcal{I}, k \in \mathcal{K} && (2.2\text{e}) \\
& D_{k,j} \geq \sqrt{(x_j - x_k)^2 + (y_j - y_k)^2} && \forall\ k \in \mathcal{K}, j \in \mathcal{K} && (2.2\text{f}) \\
& \sum_k f_{i,k} \leq a_i && \forall\ i \in \mathcal{I} && (2.2\text{g}) \\
& \sum_i f_{i,k} \cdot cv_k = f_k && \forall\ k \in \mathcal{K} && (2.2\text{h}) \\
& f_k = \sum_j f_{k,j} && \forall\ k \in \mathcal{K} && (2.2\text{i}) \\
& \sum_k f_{k,j} = d_j && \forall\ j \in \mathcal{J} && (2.2\text{j}) \\
& w_k \geq z_{i,k} && \forall\ i \in \mathcal{I}, k \in \mathcal{K} && (2.2\text{k}) \\
& \sum_i z_{i,k} \geq w_k && \forall\ k \in \mathcal{K} && (2.2\text{l}) \\
& w_k \geq z_{k,j} && \forall\ k \in \mathcal{K}, j \in \mathcal{J} && (2.2\text{m}) \\
& \sum_j z_{k,j} \geq w_k && \forall\ k \in \mathcal{K} && (2.2\text{n}) \\
& 0 \leq f_k \leq mc_k \cdot w_k && \forall\ k \in \mathcal{K} && (2.2\text{o}) \\
& 0 \leq x_k \leq x_k^{\mathrm{U}} \cdot w_k && \forall\ k \in \mathcal{K} && (2.2\text{p}) \\
& 0 \leq y_k \leq y_k^{\mathrm{U}} \cdot w_k && \forall\ k \in \mathcal{K} && (2.2\text{q}) \\
& 0 \leq f_{i,k} \leq f_{i,k}^{\mathrm{U}} \cdot z_{i,k} && \forall\ i \in \mathcal{I}, k \in \mathcal{K} && (2.2\text{r}) \\
& D_{i,k}^{\mathrm{L}} \cdot z_{i,k} \leq D_{i,k} \leq D_{i,k}^{\mathrm{U}} \cdot z_{i,k} && \forall\ i \in \mathcal{I}, k \in \mathcal{K} && (2.2\text{s}) \\
& 0 \leq f_{k,j} \leq f_{k,j}^{\mathrm{U}} \cdot z_{k,j} && \forall\ k \in \mathcal{K}, j \in \mathcal{J} && (2.2\text{t}) \\
& D_{k,j}^{\mathrm{L}} \cdot z_{k,j} \leq D_{k,j} \leq D_{k,j}^{\mathrm{U}} \cdot z_{k,j} && \forall\ k \in \mathcal{K}, j \in \mathcal{J} && (2.2\text{u}) \\
& w_k, z_{i,k}, z_{k,j} \in \{0, 1\} && \forall\ i \in \mathcal{I}, k \in \mathcal{K}, j \in \mathcal{J} && (2.2\text{v})
\end{aligned}
$$

We assume that the facilities of the same type have the same costs and characteristics associated with them, i.e., ff_k, vf_k, mc_k, $ft_{i,k}$, $vt_{i,k}$, $ft_{k,j}$, $vt_{k,j}$ are the same $\forall\ k \in \mathcal{K}_n$. Therefore, in order to avoid symmetry in the solution, we add Equations (2.2w)-(2.2x) to the formulation. These constraints enforce that for facilities k of the same type, i.e., $k \in \mathcal{K}_n,\ n \in \mathcal{N}$, the model will chose first to build the ones with the lower indices, and those will be located in lower x_k coordinate.

$$w_k \geq w_{k+1} \qquad \forall\ k \in \mathcal{K}_n,\ n \in \mathcal{N} \tag{2.2w}$$

$$x_k \geq x_{k+1} \qquad \forall\ k \in \mathcal{K}_n,\ n \in \mathcal{N} \tag{2.2x}$$

2.2 Bilevel decomposition algorithm

Although global optimization solvers perform reasonably well for small-scale instances of the nonconvex MINLP problem (2.2), their performance scales poorly due to the loose bounds of the variables in the bilinear term, thereby becoming computationally very expensive for mid to large-scale problems. For this reason, we propose a bilevel decomposition algorithm that consists of decomposing the problem into a master problem and a subproblem. The master problem is based on a relaxation of the nonconvex MINLP (2.2), which yields an MILP that predicts the selection of facilities and their links to suppliers and customers, as well as a lower bound on the cost of problem (2.1) or (2.2). The subproblem corresponds to a nonconvex NLP of reduced dimensionality that results from fixing the binary variables w_k, $z_{i,k}$, $z_{k,j}$ in the MINLP problem (2.2), according to the binary variables predicted in the MILP master problem. Figure 2.2 shows the algorithm.

2.2.1 Master problem

The non-linearity and nonconvexity of the formulation (2.2) arise from the fact that the distances are decision variables. If the coordinates for the potential facilities are fixed, the distances can be pre-computed and used as parameters in the model. In order to take advantage of this property, the master problem partitions the space into p sub-regions, which

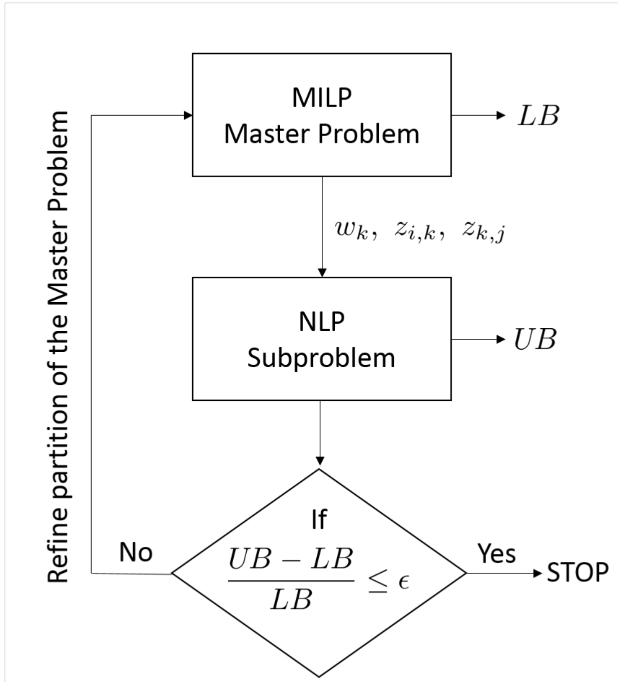

Figure 2.2: Representation bilevel decomposition algorithm

are uniform rectangular cells as represented in Figure 2.3.

In order to derive a valid relaxation for the original MINLP (2.2), we consider that each of the facilities can be located in each of the sub-regions. Therefore, it is possible to determine *a priori* the minimum distance between suppliers and facilities, and between facilities and customers. Specifically, by discretizing the 2-dimensional space, we are able to pre-calculate the minimum distance between the fixed points and each sub-region p, $\widehat{D}_{i,p}$ and $\widehat{D}_{j,p}$, as follows:

$$dx_{i,p} = \max\{|x_i - x_p| - \bar{x}/2, 0\} \qquad \forall\ i \in \mathcal{I}, p \in \mathcal{P} \tag{2.3a}$$

$$dy_{i,p} = \max\{|y_i - y_p| - \bar{y}/2, 0\} \qquad \forall\ i \in \mathcal{I}, p \in \mathcal{P} \tag{2.3b}$$

$$dx_{j,p} = \max\{|x_j - x_p| - \bar{x}/2, 0\} \qquad \forall\ j \in \mathcal{J}, p \in \mathcal{P} \tag{2.3c}$$

$$dy_{j,p} = \max\{|y_j - y_p| - \bar{y}/2, 0\} \qquad \forall\ j \in \mathcal{J}, p \in \mathcal{P} \tag{2.3d}$$

$$\widehat{D}_{i,p} = \max\{\sqrt{dx_{i,p}^2 + dy_{i,p}^2}, D_{i,p}^{\mathrm{L}}\} \qquad \forall\ i \in \mathcal{I}, p \in \mathcal{P} \tag{2.3e}$$

$$\widehat{D}_{j,p} = \max\{\sqrt{dx_{j,p}^2 + dy_{j,p}^2}, D_{j,p}^{\mathrm{L}}\} \qquad \forall\ j \in \mathcal{J}, p \in \mathcal{P} \tag{2.3f}$$

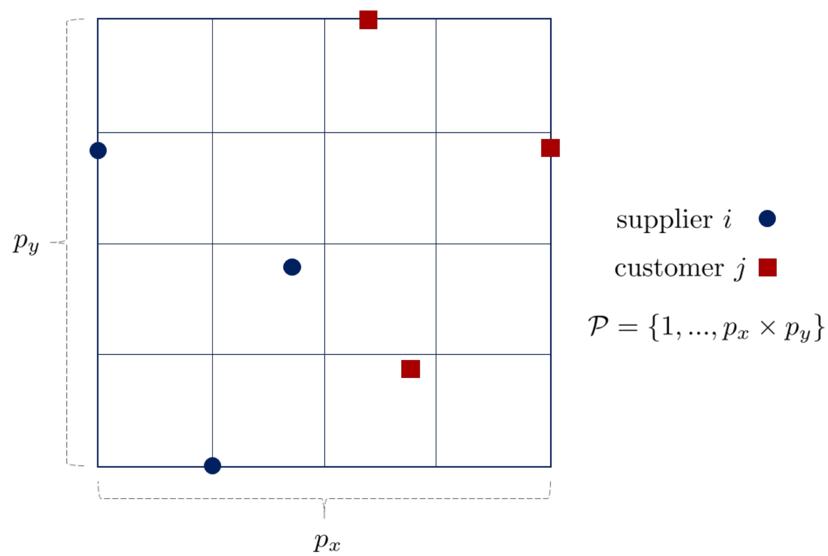

Figure 2.3: Representation of $p \in \mathcal{P}$ sub-regions

where (x_p, y_p) are the coordinates of the mid-point of each sub-region p; $\bar{x}$ and $\bar{y}$ are the length of sub-region p in the x and y directions, respectively; $D^{\mathrm{L}}_{j,p}$ and $D^{\mathrm{L}}_{i,p}$ are the lower bounds for the distances, not allowing the model to chose to build a facility k on top of a fixed point from a supplier or a customer.

Making use of the partitions and the minimum distances, (2.3e)-(2.3f), the MINLP reformulation (2.2) can then be rewritten as an MILP in (2.4). The objective of this model is to decide which facilities k to build in each sub-region p and how to allocate the raw-material and products between suppliers i, customers j and these facilities k while minimizing the total cost for the relaxed problem, Φ^P.

$$\min\ \Phi^{\mathrm{P}} = \sum_k \sum_p Cost_{k,p} + \sum_i \sum_k \sum_p Cost_{i,k,p} + \sum_k \sum_j \sum_p Cost_{k,j,p} \tag{2.4a}$$

$$\text{s.t. } Cost_{k,p} = ff_k \cdot w_{k,p} + vf_k \cdot f_{k,p} \qquad \forall\ k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4b}$$

$$Cost_{i,k,p} = cs_i \cdot f_{i,k,p} + ft_{i,k} \cdot z_{i,k,p} + vt_{i,k} \cdot \widehat{D}_{i,p} \cdot f_{i,k,p} \qquad \forall\ i \in \mathcal{I}, k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4c}$$

$$Cost_{k,j,p} = ft_{k,j} \cdot z_{k,j,p} + vt_{k,j} \cdot \widehat{D}_{j,p} \cdot f_{k,j,p} \qquad \forall\ k \in \mathcal{K}, j \in \mathcal{J}, p \in \mathcal{P} \tag{2.4d}$$

$$\sum_k \sum_p f_{i,k,p} \leq a_i \qquad \forall\ i \in \mathcal{I} \tag{2.4e}$$

$$\sum_{i} f_{i,k,p} \cdot cv_k = f_{k,p} \qquad \forall\ k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4f}$$

$$f_{k,p} = \sum_{j} f_{k,j,p} \qquad \forall\ k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4g}$$

$$\sum_{k} \sum_{p} f_{k,j,p} = d_j \qquad \forall\ j \in \mathcal{J} \tag{2.4h}$$

$$w_{k,p} \geq z_{i,k,p} \qquad \forall\ i \in \mathcal{I}, k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4i}$$

$$\sum_{i} z_{i,k,p} \geq w_{k,p} \qquad \forall\ k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4j}$$

$$w_{k,p} \geq z_{k,j,p} \qquad \forall\ k \in \mathcal{K}, j \in \mathcal{J}, p \in \mathcal{P} \tag{2.4k}$$

$$\sum_{j} z_{k,j,p} \geq w_{k,p} \qquad \forall\ k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4l}$$

$$\sum_{p} w_{k,p} \leq 1 \qquad \forall k \in \mathcal{K} \tag{2.4m}$$

$$\sum_{p} z_{i,k,p} \leq 1 \qquad \forall i \in \mathcal{I}, k \in \mathcal{K} \tag{2.4n}$$

$$\sum_{p} z_{k,j,p} \leq 1 \qquad \forall k \in \mathcal{K}, j \in \mathcal{J} \tag{2.4o}$$

$$\sum_{p' \leq p} w_{k',p'} \geq \sum_{p} w_{k,p} \qquad \forall k' < k, k \in \mathcal{K}_n, n \in \mathcal{N} \tag{2.4p}$$

$$0 \leq f_{k,p} \leq mc_k \cdot w_{k,p} \qquad \forall k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4q}$$

$$0 \leq f_{i,k,p} \leq f^{\mathrm{U}}_{i,k,p} \cdot z_{i,k,p} \qquad \forall i \in \mathcal{I}, k \in \mathcal{K}, p \in \mathcal{P} \tag{2.4r}$$

$$0 \leq f_{k,j,p} \leq f^{\mathrm{U}}_{k,j,p} \cdot z_{k,j,p} \qquad \forall k \in \mathcal{K}, j \in \mathcal{J}, p \in \mathcal{P} \tag{2.4s}$$

$$w_{k,p}, z_{i,k,p}, z_{k,j,p} \in \{0, 1\} \qquad \forall i \in \mathcal{I}, k \in \mathcal{K}, j \in \mathcal{J}, p \in \mathcal{P} \tag{2.4t}$$

Note that the MILP (2.4) has a considerably larger number of variables and constraints than the MINLP (2.2) since they must be defined for each partition p. It can easily be shown that the MILP master problem yields a lower bound to the total cost.

Proposition 1. The MILP master problem (2.4) yields a lower bound to the original MINLP problem (2.2).

Proof. First we note that although the variables are disaggregated by partitions, the inequalities in (2.4m), (2.4n), (2.4o) ensure that only one facility and one link in the original MINLP are selected. Second, since we consider the shortest distance between the fixed points and the sub-regions as in (2.3e)-(2.3f), the transportation costs are underestimated. Thus, the MILP yields a Lower Bound (LB) to the original problem, i.e., $\Phi^P \leq \Phi$. □

2.2.2 Subproblem

The subproblem consists of solving (2.2) for the fixed decisions of which facilities k to build, $\widehat{w}_k$, and how to allocate their material supply $\widehat{z}_{i,k}$ and products $\widehat{z}_{k,j}$ as selected in the MILP (2.4). The subproblem (2.5) is a reduced nonconvex NLP that is solved using a global optimization solver.

$$\min\ \Phi^{\mathrm{N}} = \sum_k Cost_k + \sum_i \sum_k Cost_{i,k} + \sum_k \sum_j Cost_{k,j} \tag{2.5a}$$

$$\text{s.t. } Cost_k = ff_k \cdot \widehat{w}_k + vf_k \cdot f_k \qquad \forall\ k \in \mathcal{K} \tag{2.5b}$$

$$Cost_{i,k} = cs_i \cdot f_{i,k} + ft_{i,k} \cdot \widehat{z}_{i,k} + vt_{i,k} \cdot f_{i,k} \cdot D_{i,k} \qquad \forall\ i \in \mathcal{I}, k \in \mathcal{K} \tag{2.5c}$$

$$Cost_{k,j} = ft_{k,j} \cdot \widehat{z}_{k,j} + vt_{k,j} \cdot f_{k,j} \cdot D_{k,j} \qquad \forall\ k \in \mathcal{K}, j \in \mathcal{J} \tag{2.5d}$$

$$D_{i,k} \geq \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2} \qquad \forall\ i \in \mathcal{I}, k \in \mathcal{K} \tag{2.5e}$$

$$D_{k,j} \geq \sqrt{(x_j - x_k)^2 + (y_j - y_k)^2} \qquad \forall\ k \in \mathcal{K}, j \in \mathcal{J} \tag{2.5f}$$

$$\sum_k f_{i,k} \leq a_i \qquad \forall\ i \in \mathcal{I} \tag{2.5g}$$

$$\sum_i f_{i,k} \cdot cv_k = f_k \qquad \forall\ k \in \mathcal{K} \tag{2.5h}$$

$$f_k = \sum_j f_{k,j} \qquad \forall\ k \in \mathcal{K} \tag{2.5i}$$

$$\sum_k f_{k,j} = d_j \qquad \forall\ j \in \mathcal{J} \tag{2.5j}$$

$$0 \leq f_k \leq mc_k \cdot \widehat{w}_k \qquad \forall\ k \in \mathcal{K} \tag{2.5k}$$

$$\underline{x}^{\mathrm{L}}_{k,p} \cdot \widehat{w}_k \leq x_k \leq \bar{x}^{\mathrm{U}}_{k,p} \cdot \widehat{w}_k \qquad \forall\ k \in \mathcal{K} \tag{2.5l}$$

$$\underline{y}^{\mathrm{L}}_{k,p} \cdot \widehat{w}_k \leq y_k \leq \bar{y}^{\mathrm{U}}_{k,p} \cdot \widehat{w}_k \qquad \forall\ k \in \mathcal{K} \tag{2.5m}$$

$$0 \leq f_{i,k} \leq f^{\mathrm{U}}_{i,k} \cdot \widehat{z}_{i,k} \qquad \forall\ i \in \mathcal{I}, k \in \mathcal{K} \tag{2.5n}$$

$$\underline{D}^{\mathrm{L}}_{i,k,p} \cdot \widehat{z}_{i,k} \leq D_{i,k} \leq \bar{D}^{\mathrm{U}}_{i,k,p} \cdot \widehat{z}_{i,k} \qquad \forall\ i \in \mathcal{I}, k \in \mathcal{K} \tag{2.5o}$$

$$0 \leq f_{k,j} \leq f^{\mathrm{U}}_{k,j} \cdot \widehat{z}_{k,j} \qquad \forall\ k \in \mathcal{K}, j \in \mathcal{J} \tag{2.5p}$$

$$\underline{D}^{\mathrm{L}}_{k,j,p} \cdot \widehat{z}_{k,j} \leq D_{k,j} \leq \bar{D}^{\mathrm{U}}_{k,j,p} \cdot \widehat{z}_{k,j} \qquad \forall\ k \in \mathcal{K}, j \in \mathcal{J} \tag{2.5q}$$

The NLP subproblem (2.5), which comprises the original problem (2.2) for a fixed set of discrete decisions, yields an Upper Bound (UB) to the total cost, $\Phi^N \geq \Phi$. The key point in the NLP is that we update the bounds of the facilities coordinates such that their location (x_k, y_k) has to be within the bounds of the sub-region p chosen in the Master Problem, i.e., for a p such that $w_{k,p} = 1$ in the solution of Problem (2.4) we have that $\underline{x}^{\mathrm{L}}_{k,p} \leq x_k \leq \bar{x}^{\mathrm{U}}_{k,p}$ and $\underline{y}^{\mathrm{L}}_{k,p} \leq y_k \leq \bar{y}^{\mathrm{U}}_{k,p}$, where:

$$\underline{x}^{\mathrm{L}}_{k,p} = x_p - \bar{x}/2 \qquad \forall\ k \in \mathcal{K} \tag{2.6a}$$

$$\bar{x}^{\mathrm{U}}_{k,p} = x_p + \bar{x}/2 \qquad \forall\ k \in \mathcal{K} \tag{2.6b}$$

$$\underline{y}^{\mathrm{L}}_{k,p} = y_p - \bar{y}/2 \qquad \forall\ k \in \mathcal{K} \tag{2.6c}$$

$$\bar{y}^{\mathrm{U}}_{k,p} = y_p + \bar{y}/2 \qquad \forall\ k \in \mathcal{K} \tag{2.6d}$$

This assumption greatly impacts tractability because the bounds for $D_{i,k}$ and $D_{k,j}$, which are part of the bilinear terms, become tighter, i.e., $\underline{D}^{\mathrm{L}}_{i,k,p} \leq D_{i,k} \leq \bar{D}^{\mathrm{U}}_{i,k,p}$ and $\underline{D}^{\mathrm{L}}_{k,j,p} \leq D_{k,j} \leq \bar{D}^{\mathrm{U}}_{k,j,p}$, where:

$$\underline{D}^{\mathrm{L}}_{i,k,p} = \widehat{D}_{i,p} \qquad \forall\ i \in \mathcal{I},\ k \in \mathcal{K} \tag{2.7a}$$

$$\bar{D}^{\mathrm{U}}_{i,k,p} = \widehat{D}_{i,p} + \sqrt{\bar{x}^2 + \bar{y}^2} \qquad \forall\ i \in \mathcal{I},\ k \in \mathcal{K} \tag{2.7b}$$

$$\underline{D}^{\mathrm{L}}_{k,j,p} = \widehat{D}_{j,p} \qquad \forall\ j \in \mathcal{J},\ k \in \mathcal{K} \tag{2.7c}$$

$$\bar{D}^{\mathrm{U}}_{k,j,p} = \widehat{D}_{j,p} + \sqrt{\bar{x}^2 + \bar{y}^2} \qquad \forall\ j \in \mathcal{J},\ k \in \mathcal{K} \tag{2.7d}$$

Thus the McCormick convex envelopes (McCormick, 1976) also become tighter, strengthening the lower bounds in the global optimization search of this NLP.

2.2.3 Algorithm

As discussed earlier in this section, the bilevel decomposition algorithm consists of iteratively solving the MILP master problem and the NLP subproblem, and refining the partitioning of space at each iteration. By shrinking the sub-regions, the minimum distances get closer to the actual distances, thus the lower bound becomes tighter; the selection of which facilities to build in each region gets closer to the optimal; and the bounds for $D_{i,k}$, and $D_{k,j}$ in the subproblem becomes tighter; hence, it is easier for the global optimization solver to find the optimal solution. The only drawback is that the size of the MILP master problem becomes larger, and thus harder to solve. The proposed procedure is described in Algorithm 1.

Algorithm 1 Bilevel decomposition algorithm for CMWP

for a pre-specified optimality tolerance of ϵ **do**
  Determine the tightest rectangle $x \times y$ that includes all the fixed points;
  Partition this rectangle into equally sized rectangles $p_x \times p_y$;
  Set iteration $iter = 1$;
  while $gap > \epsilon$ **do**
    Solve the MILP master problem (2.4) and compute Lower Bound (LB^{iter});
    Fix the decision of which facilities k to build, w_k;
    Fix the decisions of how to allocate the materials, $z_{i,k}$, and products, $z_{k,j}$;
    Fix the bounds for x_k and y_k according to the sub-region p that was selected for facility k to be built;
    Solve the nonconvex NLP subproblem (2.5) using a global optimization solver, and compute an Upper Bound (UB^{iter});
    $gap = UB^{iter} - LB^{iter}$
    $p_x = p_x + N_x$ and $p_y = p_y + N_y$, where $N_x, N_y \in \mathbb{N}$
    $iter = iter + 1$
  end while
end for

The partitioning does not need to be uniform. For a fixed number of facilities, an nonuniform grid would perform faster since only the sub-regions that had potential to place a facility would have their grid refined. Thus, the algorithm would not waste computational power in refining the sub-regions that will not have any facility. However, for the problem that is proposed in (2.1), the number of facilities k to be built is a decision variable. Based on experiments, the model tends to pick fewer and larger facilities in the early stages of the partitioning, and the structure of the network can completely change from one iteration to another. Therefore, we decided to adopt a uniform partitioning in this chapter.

A bottleneck in the performance of this procedure is the solution of the subproblem, which is nonconvex NLP. However, it does not have to be solved to global optimality to yield a valid upper bound. Therefore, we specify a maximum solution time for the subproblem so that the algorithm does not waste time trying to achieve global optimality in the early iterations, when the bounds for $D_{i,k}$, and $D_{k,j}$ are still loose.

The proposed bilevel decomposition algorithm 1 converges to the global optimum in a finite number of steps with an ϵ-tolerance. We first establish the following proposition.

Proposition 2. For an infinite number of partitions the MILP (2.4) and the MINLP (3.2) yield the same optimal solution Φ^*.

Proof. It trivially follows that for an infinite number of partitions the MILP (2.4) becomes an exact infinite dimensional representation of the MINLP (3.2). Thus, both (2.2) and (2.4) have the same optimal solution Φ^*. $\square$

Theorem 1. Algorithm 1 converges in a finite number of iterations to the global optimum of problem (2.2) within an ϵ-tolerance at the bounds, $LB \geq UB - \epsilon$.

Proof. Algorithm 1 consists at solving a sequence of MILP problems (2.4) by increasing the number of partitions such that the set of partitions $\mathcal{P}^{iter}$ at iteration $iter$ is contained in the next iteration $\mathcal{P}^{iter} \subset \mathcal{P}^{iter+1}$. Thus, it follows that the lower bounds from (2.4) satisfy $\Phi^{P,iter} \leq \Phi^{P,iter+1}$. From Proposition 2 we have that $\Phi^{P,\infty} = \Phi^*$. Since we only consider a finite number of iterations, it follows that $\Phi^{P,iter} < \Phi^*$. Since the NLP subproblem (2.5) is solved to global optimality, it follows that $\Phi^{P,iter} < \Phi^* \leq \Phi^{\widehat{N}}$, where $\Phi^{\widehat{N}}$ is the incumbent, i.e. the best feasible solution of the NLP subproblem (2.5). For a given tolerance ϵ, a finite number of iterations $\widehat{iter}$ can be selected such that $\Phi^{P,\widehat{iter}} \geq \Phi^{\widehat{N}} - \epsilon$. Since $\Phi^{P,\widehat{iter}} = LB$, and $\Phi^{\widehat{N}} = UB$, $LB \geq UB - \epsilon$. Thus, the algorithm converges with an ϵ tolerance in a finite number of steps. $\square$

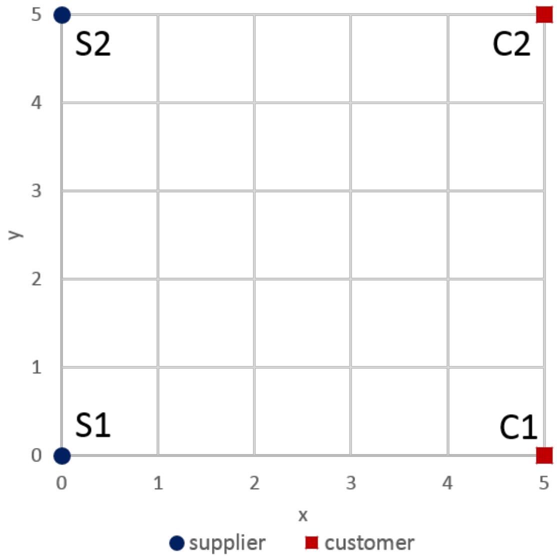

Figure 2.4: Small test problem's network

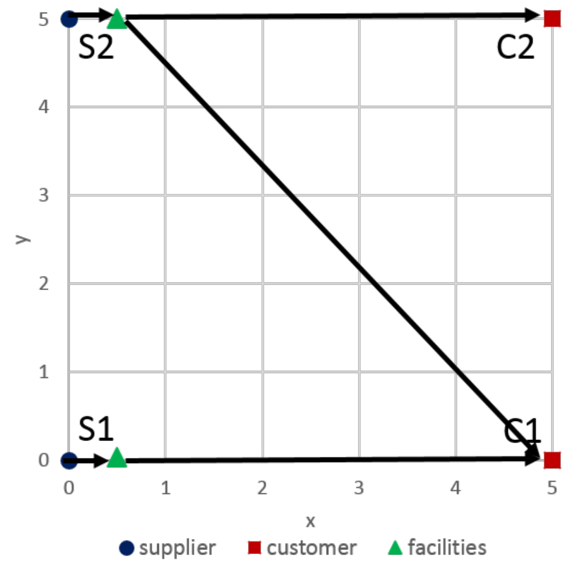

Figure 2.5: Small test problem's optimal network

2.3 Small Test Problem

In order to test the Algorithm 1, we applied it first to a small test problem with 2 suppliers and 2 customer points. Supplier 1 is located at coordinates (0,0), and supplier 2 is located at (0,5). Customer 1 is located at coordinates (5,0), and customer 2 is located at (5,5). The fixed points of the network are represented in Figure 2.4. The cost of supply material from supplier 1 is 20 and from supplier 2 is 22. Both suppliers have an availability of 120, and both markets have a demand of 100.

There are 2 types of facilities. Type 1 has two potential facilities with a maximum capacity of 125 each, fixed cost of 7.18, and variable cost of 0.087. Type 2 has one potential facility with a maximum capacity of 250, fixed cost of 10.77, and variable cost of 0.067. All the facilities have a conversion to product flow of 90%.

The fixed transportation costs, $ft_{i,k}$ and $ft_{i,k}$, are 10, and the variable transportation costs, $vt_{i,k}$ and $vt_{i,k}$, are 0.3 for any type of link. It is assumed that the minimum allowed distance between a fixed supply or customer point and a facility is 0.5. The MILP master problem is solved to 0.01% optimality gap, and the maximum CPU time allowed for the NLP subproblem is 200 seconds. The monolithic MINLP version of the problem, (3.2), has

Table 2.1: Small test problem results

	Lower Bound	Upper Bound	Gap
$iter = 1$	4776.392	5159.830	8.028%
$iter = 2$	4916.468	5039.304	2.498%
$iter = 3$	4946.704	5039.304	1.872%
$iter = 4$	4968.799	5039.304	1.419%
$iter = 5$	4982.116	5039.304	1.148%
$iter = 6$	4991.011	5039.304	0.968%
$iter = 7$	4997.371	5039.304	0.839%
$iter = 8$	5002.145	5039.304	0.743%
$iter = 9$	5005.859	5039.304	0.668%
$iter = 10$	5008.832	5039.304	0.608%
$iter = 11$	5011.265	5039.304	0.560%
$iter = 12$	5013.293	5039.304	0.519%
$iter = 13$	5015.009	5039.304	0.484%

68 constraints, 15 binary variables, and 38 continuous variables.

Starting from p_x and p_y equal 1, i.e., no partition, and increasing them by 1 at each iteration, it takes 13 iterations and 11.87 seconds to solve this problem to 0.5% optimality tolerance. The lower bound, upper bound and optimality gap at each iteration are reported in Table 2.1.

As one can see, the lower bound gradually tightens up as the number of iterations $iter$, and consequently the number of partitions increase. The optimal solution of 5039.304 is the same as found by the general purpose global optimization solvers. However, while the algorithm solved this problem within an optimality tolerance of 0.5% in 11.87 seconds, BARON solved it in 1247.10 seconds, and ANTIGONE and SCIP could not solve it in 1 hour. The optimal network is shown in Figure 2.5.

2.4 Computational Results

In order to compare the performance of our proposed algorithm with currently available general purpose global optimization solvers, we randomly generated 15 test cases. The network varies in size as follows.

- Network 1: 2 suppliers $\times$ 2 consumers;
- Network 2: 5 suppliers $\times$ 5 consumers;
- Network 3: 10 suppliers $\times$ 10 consumers;
- Network 4: 20 suppliers $\times$ 20 consumers;
- Network 5: 40 suppliers $\times$ 40 consumers;

The 5 network structures are represented in Figures 2.6-2.10.

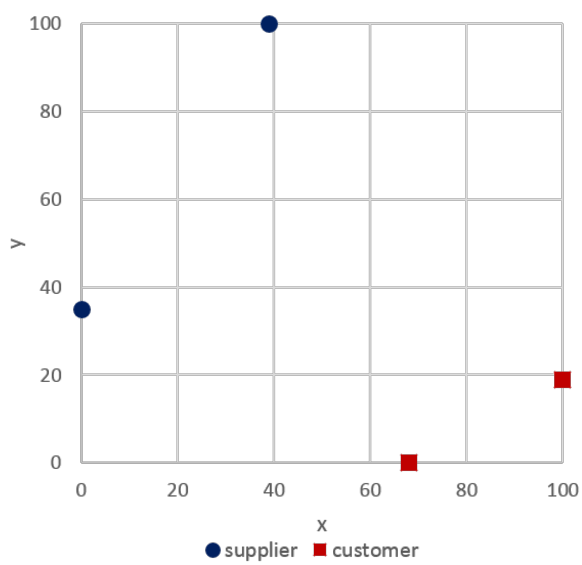

Figure 2.6: Network 1

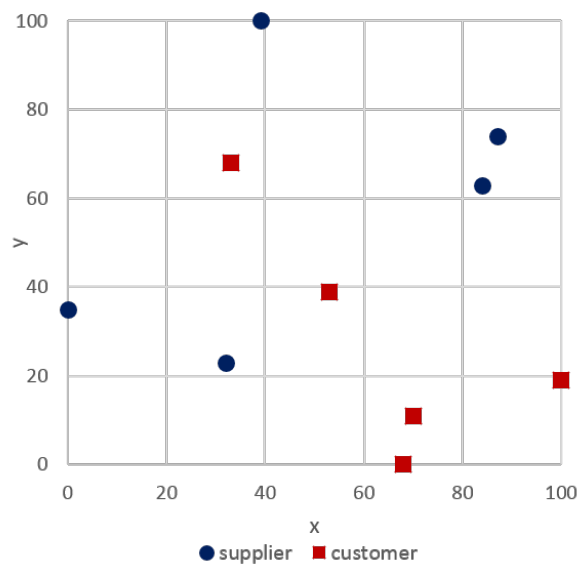

Figure 2.7: Network 2

For each of the network options, the choice of 1, 2 and 3 types of facilities were tested, such that:

- Type 1: up to 2 large-scale facilities
- Type 2: up to 10 mid-scale facilities;
- Type 3: up to 20 small-scale facilities;

Therefore, for each of the network structure the problem was solved for 2, 12 and 32 potential facilities.

Each test case was solved using Algorithm 1 and by general purpose global optimization solvers, BARON, ANTIGONE and SCIP. We set the optimality tolerance to 1% and the maximum total CPU time to 1 hour. Regarding the algorithm, it is required that the master

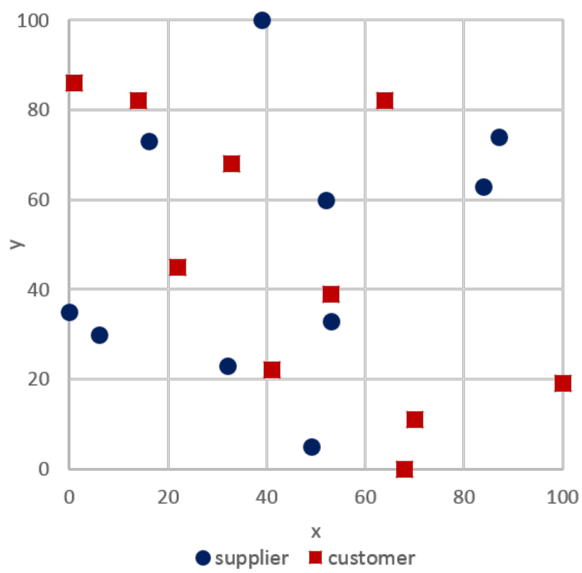

Figure 2.8: Network 3

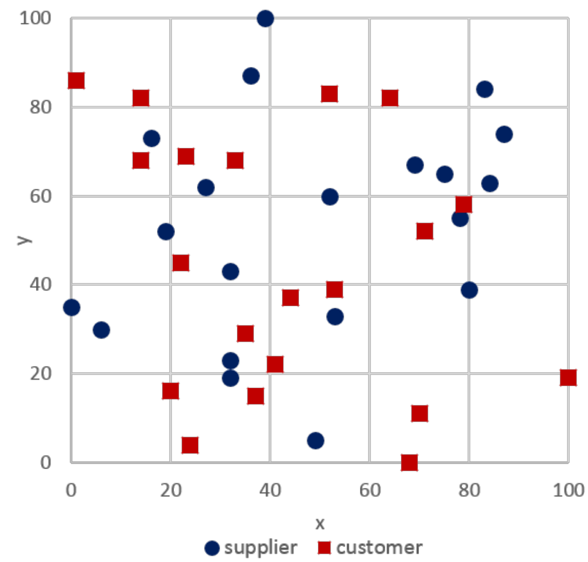

Figure 2.9: Network 4

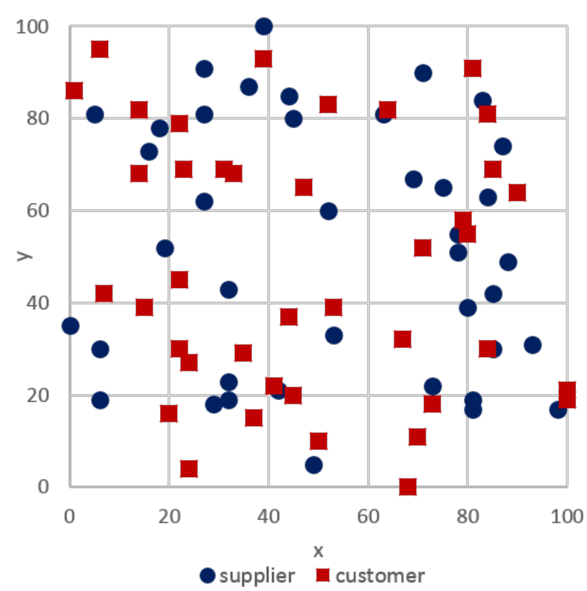

Figure 2.10: Network 5

problem has to be solved to 0.1% optimality gap, and it is allowed a maximum CPU time of 200 seconds for the solution of each NLP subproblem. We start the algorithm with a 10×10 partitioning of the space and at each iteration this partitioning is increased by $N_x, N_y = 5$.

Our computational tests were performed on a standard desktop computer with an Intel(R)

Core(TM) i7-2600 CPU @ 3.40 GHz processor, with 8GB of RAM, running on Windows 7. We implemented the monolithic formulation and the global optimization algorithm in GAMS 24.7.1, solve the MILPs using CPLEX version 12.6.3 (IBM, 2015), the NLPs using BARON version 16.3.4 (Tawarmalani and Sahinidis, 2005), and the MINLPs using BARON version 16.3.4 (Tawarmalani and Sahinidis, 2005), ANTIGONE (Misener and Floudas, 2014), and SCIP (Gamrath et al., 2016).

The case-studies are named such that the first 2 letters represent the network (i.e., N1, N2, N3, N4, and N5, represent Network 1, 2, 3, 4, and 5, respectively), and the last 2 letters represent the number of facility types considered (i.e., T1, T2, T3 represent 1 type, 2 types and 3 types, respectively). The size of monolithic MINLP formulation (3.2) for each of the test cases is shown in Table 2.2, and their results are shown in Table 2.3.

Table 2.2: Monolithic MINLP formulation size

	Binary Variables	Continuous Variables	Constraints
N1-T1	10	27	49
N2-T1	22	51	91
N3-T1	42	91	161
N4-T1	82	171	301
N5-T1	162	331	581
N1-T2	197	137	329
N2-T2	132	281	551
N3-T2	252	521	921
N4-T2	492	1001	1661
N5-T2	972	1961	3141
N1-T3	160	357	1089
N2-T3	352	741	1671
N3-T3	672	1381	2641
N4-T3	1312	2661	4581
N5-T3	2592	5221	8461

The results in Table 2.3 show that the bilevel decomposition algorithm was able to find the optimal solution within 1% optimality tolerance in 87% of the the case studies, and performed better than the other general purpose optimization solvers in 73% of them. It can be noticed that the improvement in performance due to the use of the algorithm is clearer for

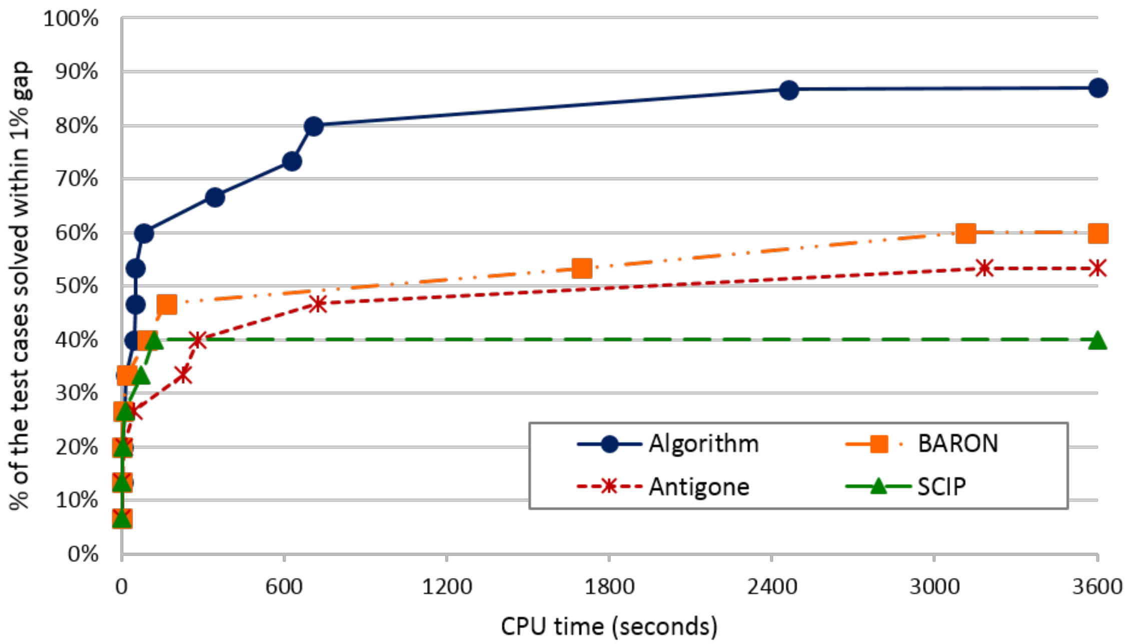

Figure 2.11: Performance curves comparing the bilevel decomposition algorithm with global optimization solvers

larger instances, specifically the networks with larger number of supplier and customer fixed points. The global optimization solver that performed the best for this type of problem was BARON. Antigone was the global solver that had the worst performance, not being able to find a feasible solution in 47% of the test cases. The performance curves for the Algorithm 1 and each of the global solvers are shown in Figure 2.11 (Dolan and Moré, 2002).

Table 2.3: Computational experiments' results

		Algorithm 1	BARON	ANTIGONE	SCIP
	Minimum Cost	10.537	10.537	10.537	10.537
N1-T1	Gap	0.4%	1.0%	1.0%	1.0%
	CPU Time (s)	0.725	0.090	0.267	0.360
	Minimum Cost	23.850	23.850	23.850	23.850
N2-T1	Gap	0.8%	1.0%	1.0%	1.0%
	CPU Time (s)	5.046	2.480	8.014	12.780

N3-T1	Minimum Cost	44.913	44.957	44.913	44.932
	Gap	0.9%	1.0%	1.0%	3.3%
	CPU Time (s)	45.854	164.330	225.297	3600
N4-T1	Minimum Cost	60.410	60.416	60.617	60.491
	Gap	0.8%	1.6%	1.0%	4.18%
	CPU Time (s)	340.541	3600	3183.016	3600
N5-T1	Minimum Cost	91.926	92.515	Infeasible	93.253
	Gap	0.8%	3.5%	NA	8.4%
	CPU Time (s)	705.269	3600	0.311	3600
N1-T2	Minimum Cost	10.537	10.537	10.537	10.537
	Gap	0.4%	1.0%	1.0%	1.0%
	CPU Time (s)	3.243	1.910	1.093	2.040
N2-T2	Minimum Cost	23.850	23.850	23.850	23.850
	Gap	0.8%	1.0%	1.0%	1.0%
	CPU Time (s)	12.055	18.810	280.717	73.980
N3-T2	Minimum Cost	44.913	44.913	No solution returned	44.989
	Gap	0.9%	1.0%	NA	3.6%
	CPU Time (s)	50.038	3114.320	3600	3600
N4-T2	Minimum Cost	60.411	60.425	No solution returned	61.042
	Gap	0.8%	2.0%	NA	5.73%
	CPU Time (s)	627.567	3600	3600	3600
N5-T2	Minimum Cost	91.966	92.466	Infeasible	94.362
	Gap	1.2%	5.5%	NA	10.2%
	CPU Time (s)	3600	3600	7.491	3600
N1-T3	Minimum Cost	3.921	3.921	3.921	3.921
	Gap	0.6%	1.0%	1.0%	1.0%
	CPU Time (s)	15.520	3.921	47.262	3.480

N2-T3	Minimum Cost	23.850	23.850	23.850	23.850
	Gap	0.8%	1.0%	1.0%	1.0%
	CPU Time (s)	51.521	93.350	726.475	118.430
N3-T3	Minimum Cost	44.932	44.913	No solution returned	44.989
	Gap	0.9%	1.0%	NA	3.46%
	CPU Time (s)	80.239	1697.660	3600	3600
N4-T3	Minimum Cost	60.408	60.459	No solution returned	61.235
	Gap	0.8%	2.1%	NA	5.7%
	CPU Time (s)	2461.583	3600	3600	3600
N5-T3	Minimum Cost	92.621	92.2545	Infeasible	94.2516
	Gap	2.0%	5.4%	NA	10.1%
	CPU Time (s)	3600	3600	57.453	3600

2.5 Conclusions

In this chapter we have presented a new version of the Capacitated Multi-facility Weber Problem that has fixed costs, multiple types of facilities, and two sets of fixed points representing suppliers and consumers. We have proposed a GDP formulation for this problem, which was reformulated as an MINLP, and then introduced a bilevel decomposition algorithm for this nonconvex problem. We prove that this algorithm converges to the global optimum within an ϵ tolerance in a finite number of iterations.

We test the algorithm for 15 test cases varying from 2 suppliers and 2 consumers, to 40 suppliers and 40 consumers, from 1 to 3 types of facilities, and from 2 to 32 potential facilities, and compare the results with general purpose global optimization solvers. The results show that our algorithm performs more efficiently for 73% of the test cases within 1% of optimality gap, and that the improvement in performance is more noticeable for larger instances.

Chapter 3

Global Optimization Algorithm for Multi-period Design and Planning of Centralized and Distributed Manufacturing Networks

In this chapter, we extend the work of chapter 2 to solve the design and multi-period planning of centralized and distributed manufacturing networks. The model proposed in this chapter is a multi-period nonlinear Generalized Disjunctive Programming (GDP), reformulated as a multi-period nonconvex Mixed-Integer Nonlinear Programming (MINLP). Due to the extra layer of complexity added by the multi-period formulation, we propose an accelerated version of the algorithm proposed in chapter 2 to improve its computational performance and scalability. Accordingly, the contributions of this work are on the formulation (multi-period), application (centralized and distributed networks), and additional steps to the bilevel decomposition algorithm.

The remainder of the chapter is organized as follows. We begin by presenting in Section 3.1 the problem statement. Section 3.2 includes the General Disjunctive Programming (GDP) formulation and its reformulation as a nonconvex MINLP. In Section 3.3 we propose an accelerated version of the global optimization algorithm by Lara et al. (2018c), which is guaranteed to have ϵ-convergence, and illustrate the method for a test problem. In Section 3.4 we benchmark the performance of the accelerated algorithm against the original and the commercial global solvers available for the set of randomly generated instances from Lara et al. (2018c) extended to multi-period problems. Finally, in Section 3.5 we apply the formulation and solution strategy to a biomass supply chain case study, and in Section 3.6 we draw the conclusions.

3.1 Problem Statement

Given is a set of suppliers $i \in \mathcal{I}$, with their respective fixed location coordinates (X_i, Y_i), availability $AV_{i,t}$, and cost of material supply $CRM_{i,t}$ at each time period $t \in \mathcal{T}$. Given is also a set of customers $j \in \mathcal{J}$, with their respective fixed locations (X_j, Y_j), and demands $DM_{j,t}$ per time period t. Given are the fixed and variable investment costs ($FIC_{k,t}$ and $VIC_{k,t}$, respectively) and variable operating costs ($VOC_{k,t}$) of potential facilities $k \in \mathcal{K}$ with N different types (i.e. centralized and distributed $N = 2$), which are partitioned into subsets $\mathcal{K}_n \ \forall \ n \in \mathcal{N} = \{1, ..., N\}$ such that $\bigcup_{n \in \mathcal{N}} \mathcal{K}_n = \mathcal{K}$ and $\mathcal{K}_{n_l} \cap \mathcal{K}_{n_m} = \emptyset \ \forall n_l, n_m \in \mathcal{N}, l \neq m$. The corresponding maximum capacity, MC_k, and conversion to product flows, CV_k, of these potential facilities are also known. Given are also the transportation costs between suppliers and facilities, and facilities and markets ($FTC^{\text{s}}_{i,k}, FTC^{\text{c}}_{k,j}$: fixed costs; $VTC^{\text{s}}_{i,k}, VTC^{\text{c}}_{k,j}$: variable costs). The problem is to find the optimal network of facilities (number, types, location, when to build, and corresponding flows) that minimizes the total cost.

The variables in the problem are the coordinates of potential facilities, (x_k, y_k), the distances between supplier and facility, $d^{\text{s}}_{i,k}$, and between facility and customer, $d^{\text{c}}_{k,j}$, the flows between supplier and facility, $ff^{s}_{i,k,t}$, and between facility and customer, $ff^{\text{c}}_{k,j,t}$, and the amount produced by each facility, $f_{k,t}$, in each time period t. There are also Boolean variables: $B_{k,t}$ (true if facility is built in time period t; false otherwise); $W_{k,t}$ (true if facility is in operation in time period t; false otherwise); $Z^{\text{s}}_{i,k,t}$ (true if material supply is transported between supplier and facility during time period t; false otherwise); and $Z^{c}_{k,j,t}$ (true if product is transported between facility and customer during time period t; false otherwise).

3.2 Model Formulation

3.2.1 Generalized Disjunctive Programming (GDP)

We first formulate the problem as Generalized Disjunctive Programming (GDP) to take advantage of the disjunctive structure of some of the decisions. Extending the model proposed by Lara et al. (2018c), the GDP formulation is given by Equations (3.1a)-(3.1t).

$$\min\ \Phi = \sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t} \cdot \sum_{k\in\mathcal{K}} \left(inv_{k,t} + op_{k,t} + \sum_{i\in\mathcal{I}} cost^{s}_{i,k,t} + \sum_{j\in\mathcal{J}} cost^{c}_{k,j,t} \right) \tag{3.1a}$$

$$\text{s.t.} \begin{bmatrix} B_{k,t} \\ inv_{k,t} = FIC_{k,t} + VIC_{k,t}\cdot MC_k \end{bmatrix} \vee \begin{bmatrix} \neg B_{k,t} \\ inv_{k,t} = 0 \end{bmatrix} \qquad \forall\ k\in\mathcal{K}, t\in\mathcal{T} \tag{3.1b}$$

$$\begin{bmatrix} W_{k,t} \\ op_{k,t} = VOC_{k,t}\cdot f_{k,t} \\ 0 \le f_{k,t} \le MC_k \end{bmatrix} \vee \begin{bmatrix} \neg W_{k,t} \\ op_{k,t} = 0 \\ f_{k,t} = 0 \end{bmatrix} \qquad \forall\ k\in\mathcal{K}, t\in\mathcal{T} \tag{3.1c}$$

$$\begin{bmatrix} Z^{s}_{i,k,t} \\ cost^{s}_{i,k,t} = CS_{i,t}\cdot ff^{s}_{i,k,t} + FTC^{s}_{i,k} + VTC^{s}_{i,k}\cdot ff^{s}_{i,k,t}\cdot d^{s}_{i,k} \\ 0 \le ff^{s}_{i,k,t} \le \overline{FF^{s}_{i,k,t}} \end{bmatrix} \vee \begin{bmatrix} \neg Z^{s}_{i,k,t} \\ cost^{s}_{i,k,t} = 0 \\ ff^{s}_{i,k,t} = 0 \end{bmatrix} \qquad \forall\ i\in\mathcal{I}, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.1d}$$

$$\begin{bmatrix} Z^{c}_{k,j,t} \\ cost^{c}_{k,j,t} = FTC^{c}_{k,j} + VTC^{c}_{k,j}\cdot ff^{c}_{k,j,t}\cdot d^{c}_{k,j} \\ 0 \le ff^{c}_{k,j,t} \le \overline{FF^{c}_{k,j,t}} \end{bmatrix} \vee \begin{bmatrix} \neg Z^{c}_{k,j,t} \\ cost^{c}_{k,j,t} = 0 \\ ff^{c}_{k,j,t} = 0 \end{bmatrix} \qquad \forall\ k\in\mathcal{K}, j\in\mathcal{J}, t\in\mathcal{T} \tag{3.1e}$$

$$\begin{bmatrix} \bigvee_{t\in\mathcal{T}} B_{k,t} \\ 0 \le x_k \le \overline{X_k} \\ 0 \le y_k \le \overline{Y_k} \end{bmatrix} \vee \begin{bmatrix} \neg \bigvee_{t\in\mathcal{T}} B_{k,t} \\ x_k = 0 \\ y_k = 0 \end{bmatrix} \qquad \forall\ k\in\mathcal{K} \tag{3.1f}$$

$$d^{s}_{i,k} \ge \sqrt{(X_i - x_k)^2 + (Y_i - y_k)^2} \qquad \forall\ i\in\mathcal{I}, k\in\mathcal{K} \tag{3.1g}$$

$$d^{c}_{k,j} \ge \sqrt{(X_j - x_k)^2 + (Y_j - y_k)^2} \qquad \forall\ k\in\mathcal{K}, j\in\mathcal{J} \tag{3.1h}$$

$$W_{k,t} \iff \bigvee_{i\in\mathcal{I}} Z^{s}_{i,k,t} \qquad \forall\ k\in\mathcal{K}, t\in\mathcal{T} \tag{3.1i}$$

$$W_{k,t} \iff \bigvee_{j\in\mathcal{J}} Z^{c}_{k,j,t} \qquad \forall\ k\in\mathcal{K}, t\in\mathcal{T} \tag{3.1j}$$

$$W_{k,t} \iff W_{k,t-1} \vee B_{k,t} \qquad \forall\ k\in\mathcal{K} \tag{3.1k}$$

$$\sum_{k\in\mathcal{K}} f^{s}_{i,k,t} \le AV_{i,t} \qquad \forall\ i\in\mathcal{I}, t\in\mathcal{T} \tag{3.1l}$$

$$\sum_{i\in\mathcal{I}} f^{s}_{i,k,t}\cdot CV_k = f_{k,t} \qquad \forall\ k\in\mathcal{K}, t\in\mathcal{T} \tag{3.1m}$$

$$f_{k,t} = \sum_{j \in \mathcal{J}} f^{c}_{k,j,t} \qquad \forall\ k \in \mathcal{K}, t \in \mathcal{T} \tag{3.1n}$$

$$\sum_{k \in \mathcal{K}} f^{c}_{k,j,t} = DM_{j,t} \qquad \forall\ j \in \mathcal{J}, t \in \mathcal{T} \tag{3.1o}$$

$$w_k \geq w_{k+1} \qquad \forall\ k \in \mathcal{K}_n,\ n \in \mathcal{N} \tag{3.1p}$$

$$x_k \geq x_{k+1} \qquad \forall\ k \in \mathcal{K}_n,\ n \in \mathcal{N} \tag{3.1q}$$

$$D^{min} \leq d^{s}_{i,k} \leq D^{max} \qquad \forall\ k \in \mathcal{K} \tag{3.1r}$$

$$D^{min} \leq d^{c}_{k,j} \leq D^{max} \qquad \forall\ k \in \mathcal{K} \tag{3.1s}$$

$$B_{k,t}, W_{k,t}, Z^{s}_{i,k,t}, Z^{c}_{k,j,t} \in \{True, False\} \qquad \forall\ i \in \mathcal{I}, k \in \mathcal{K}, j \in \mathcal{J}, t \in \mathcal{T} \tag{3.1t}$$

The objective function (3.1a) is the net present cost, which includes investment and operating costs for building and operating the facilities, and transportation cost from the suppliers and to the customers with an interest rate, R. This is different than the GDP proposed by Lara et al. (2018c) as it now includes a series of cash flows occurring at each time period, and the facility costs are divided into investment and operating costs.

Disjunction (3.1b) determines whether facility k is built at time t ($B_{k,t}$), and disjunction (3.1c) determines whether facility k is in operation at time t ($W_{k,t}$). These disjunctions indirectly address the choice between centralized and distributed facilities as each of the potential facilities have a specified type (i.e. distributed or centralized) and their characteristics and costs are drawn from their type. This differs from the formulation by Lara et al. (2018c) where multiple types are allowed instead of only two.

Disjunctions (3.1d) and (3.1e) decide if there is material flow between the transportation links $\{i, k\}$ and $\{k, j\}$ at each time period t, which is determined by the corresponding Boolean variables ($Z^{s}_{i,k,t}$, $Z^{c}_{k,j,t}$). The last disjunction, (3.1f), specifies that if a facility k is built at any point within the planning horizon, its coordinates should be within the appropriate bounds. However, if this facility is not built, then its coordinates should be set to $(0, 0)$, to avoid degeneracy in the solution. These five proposed disjunctions, (3.1b)-(3.1f), are similar to the disjunctions in the GDP model by Lara et al. (2018c), but have the additional flexibility of allowing different allocations by time-period, specifying in which time period a facility is built, and only accounting for operating costs in the time periods the facility is in operation.

Constraints (3.1g) and (3.1h) represent the Euclidean distances between suppliers and facilities, and facilities and customers, which is the same distance representation used by Lara et al. (2018c). The logic relations in (3.1i), (3.1j) and (3.1i) establish the existence of links depending on the choice of the facilities and vice-versa, and specify that a facility k can only operate ($W_{k,t}$) if it has been built before ($B_{k,t}$). Constraints (3.1l)-(3.1o) define the mass balances, as well as the availability and demands, same as in Lara et al. (2018c).

We assume that the facilities of the same type have the same costs and characteristics associated with them, i.e., $FIC_{k,t}$, $VIC_{k,t}$, $VOC_{k,t}$, $FTC_{i,k,t}$, $VTC_{i,k,t}$, $FTC_{k,j,t}$, $VTC_{k,j,t}$, MC_k, and CV_k are the same $\forall\ k \in \mathcal{K}_n$. Analogously to Lara et al. (2018c), we have constraints (3.1p)-(3.1q) to break the symmetry in the facility selection within the same type and avoid degeneracy in the solution. These constraints enforce that for facilities k of the same type, i.e., $k \in \mathcal{K}_n,\ n \in \mathcal{N}$, the model will chose first to build the ones with the lower indices, and those will be located in lower x_k coordinate. Finally, constraints (3.1r)-(3.1s) determine the bounds for the distances, D^{min} and D^{max}, and (3.1t) defined the Boolean variables.

The GDP model (3.1) is nonconvex due to the bilinear terms ($ff \cdot d$) in the transportation cost, as can be seen in disjunctions (3.1d)-(3.1e). There is a large body of literature on relaxations and reformulations of bilinear terms, most of them deriving from the McCormick envelope (McCormick, 1976): e.g. Bergamini et al. (2005); Gounaris et al. (2009); Vielma and Nemhauser (2011); Misener et al. (2011); Kolodziej et al. (2013). The presence of bilinear terms, which can give rise to local minima, is the main motivation behind choosing a GDP formulation. By having the bilinear terms as part of the disjunctions, they are calculated only for the selected connections within an iterative procedure. Therefore, for a fixed choice of Boolean variables, the GDP leads to a reduction in the number of bilinear terms and generates a more favorable structure that can be exploited in a decomposition scheme. Additionally, equations (3.1g)- (3.1h) are nonlinear convex constraints since they correspond to Euclidean norms (Al-Loughani, 1997).

3.2.2 Mixed-integer nonlinear Programming (MINLP) model

The GDP can be transformed into an MINLP using the hull reformulation, which yields the tightest relaxation for each disjunction (Trespalacios and Grossmann, 2014). Since the disaggregated variables can be reformulated back to the original variables, the resulting MINLP is given by Equations (3.2a)-(3.2x). Again, the main difference between the MINLP reformulation presented in Lara et al. (2018c) and the following MINLP is the added flexibility of allowing multi-period operating and allocation decisions, as well as accounting for operating costs by time-period.

$$\min \ \Phi = \sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t} \cdot \sum_{k\in\mathcal{K}} \left(inv_{k,t} + op_{k,t} + \sum_{i\in\mathcal{I}} cost^{\mathrm{s}}_{i,k,t} + \sum_{j\in\mathcal{J}} cost^{\mathrm{c}}_{k,j,t} \right) \tag{3.2a}$$

$$\text{s.t} \quad inv_{k,t} = (FIC_{k,t} + VIC_{k,t} \cdot MC_k) \cdot b_{k,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2b}$$

$$op_{k,t} = VOC_{k,t} \cdot f_{k,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2c}$$

$$cost^{\mathrm{s}}_{i,k,t} = CS_{i,t} \cdot f^{\mathrm{s}}_{i,k,t} + FTC^{\mathrm{s}}_{i,k} \cdot z^{\mathrm{s}}_{i,k,t} + VTC^{\mathrm{s}}_{i,k} \cdot ff^{\mathrm{s}}_{i,k,t} \cdot d^{\mathrm{s}}_{i,k} \qquad \forall\, i\in\mathcal{I}, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2d}$$

$$cost^{\mathrm{c}}_{k,j,t} = FTC^{\mathrm{c}}_{k,j} \cdot z^{\mathrm{c}}_{k,j,t} + VTC^{\mathrm{c}}_{k,j} \cdot ff^{\mathrm{c}}_{k,j,t} \cdot d^{\mathrm{c}}_{k,j} \qquad \forall\, k\in\mathcal{K}, j\in\mathcal{J}, t\in\mathcal{T} \tag{3.2e}$$

$$d^{\mathrm{s}}_{i,k} \geq \sqrt{(X_i - x_k)^2 + (Y_i - y_k)^2} \qquad \forall\, i\in\mathcal{I}, k\in\mathcal{K} \tag{3.2f}$$

$$d^{\mathrm{c}}_{k,j} \geq \sqrt{(X_j - x_k)^2 + (Y_j - y_k)^2} \qquad \forall\, k\in\mathcal{K}, j\in\mathcal{K} \tag{3.2g}$$

$$\sum_{k\in\mathcal{K}} ff^{\mathrm{s}}_{i,k,t} \leq AV_{i,t} \qquad \forall\, i\in\mathcal{I}, t\in\mathcal{T} \tag{3.2h}$$

$$\sum_{i\in\mathcal{I}} ff^{\mathrm{s}}_{i,k,t} \cdot CV_k = f_{k,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2i}$$

$$f_{k,t} = \sum_{j\in\mathcal{J}} ff^{\mathrm{c}}_{k,j,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2j}$$

$$\sum_{k\in\mathcal{K}} ff^{\mathrm{c}}_{k,j,t} = DM_{j,t} \qquad \forall\, j\in\mathcal{J}, t\in\mathcal{T} \tag{3.2k}$$

$$w_{k,t} = \sum_{i\in\mathcal{I}} z^{\mathrm{s}}_{i,k,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2l}$$

$$w_{k,t} = \sum_{j\in\mathcal{J}} z^{\mathrm{c}}_{k,j,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2m}$$

$$w_{k,t} = w_{k,t-1} + b_{k,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2n}$$

$$0 \leq f_{k,t} \leq MC_k \cdot w_{k,t} \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2o}$$

$$0 \leq ff^{\mathrm{s}}_{i,k,t} \leq \overline{FF^{\mathrm{s}}_{i,k,t}} \cdot z^{\mathrm{s}}_{i,k,t} \qquad \forall\, i\in\mathcal{I}, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.2p}$$

$$0 \leq ff^{\mathrm{c}}_{k,j,t} \leq \overline{FF^{\mathrm{c}}_{k,j,t}} \cdot z^{\mathrm{c}}_{k,j,t} \qquad \forall\, k\in\mathcal{K}, j\in\mathcal{J}, t\in\mathcal{T} \tag{3.2q}$$

$$0 \leq x_k \leq \overline{X_k} \cdot \sum_{t \in \mathcal{T}} b_{k,t} \qquad \forall\, k \in \mathcal{K} \tag{3.2r}$$

$$0 \leq y_k \leq \overline{Y_k} \cdot \sum_{t \in \mathcal{T}} b_{k,t} \qquad \forall\, k \in \mathcal{K} \tag{3.2s}$$

$$w_{k,t} \geq w_{k+1,t} \qquad \forall\, k \in \mathcal{K}_n,\; n \in \mathcal{N}, t \in \mathcal{T} \tag{3.2t}$$

$$x_k \geq x_{k+1} \qquad \forall\, k \in \mathcal{K}_n,\; n \in \mathcal{N} \tag{3.2u}$$

$$D^{min} \leq d^{\mathrm{s}}_{i,k} \leq D^{max} \qquad \forall\, i \in \mathcal{I}, k \in \mathcal{K} \tag{3.2v}$$

$$D^{min} \leq d^{\mathrm{c}}_{k,j} \leq D^{max} \qquad \forall\, k \in \mathcal{K}, j \in \mathcal{J} \tag{3.2w}$$

$$b_{k,t}, w_{k,t}, z^{\mathrm{s}}_{i,k,t}, z^{\mathrm{c}}_{k,j,t} \in \{0,1\} \qquad \forall\, i \in \mathcal{I}, k \in \mathcal{K}, j \in \mathcal{J}, t \in \mathcal{T}. \tag{3.2x}$$

The MINLP model (3.2) can be more concisely represented by (3.3).

$$\Phi = \quad \min \quad g(f, ff, z) + d^{\intercal} C ff \tag{3.3a}$$

$$\text{s.t.} \quad d_{l,k} \geq \sqrt{(X_l - x_k)^2 + (Y_l - y_k)^2} \qquad \forall l \in \mathcal{I} \cup \mathcal{J},\; k \in \mathcal{K} \tag{3.3b}$$

$$f, ff, z, d, x, y \in \Omega, \tag{3.3c}$$

where ff is the vector of all flows between suppliers and facilities, $ff^{s}_{i,k,t}$, and between facilities and customers, $ff^{\mathrm{c}}_{k,j,t}$; f is the vector of all facilities' productions at each time period, $f_{k,t}$; z is the vector of all discrete decision variables ($b_{k,t}$, $w_{k,t}$, $z^{\mathrm{s}}_{i,k,t}$, and $z^{\mathrm{c}}_{k,j,t}$); and $g(f, ff, z)$ is the cost function associated with these decision variables. Additionally, d is the vector of distances, C is the matrix of variable transportation costs ($VTC^{\mathrm{s}}_{i,k}$ and $VTC^{\mathrm{c}}_{k,j}$), and ($d^{\intercal} C ff$) is the bilinear term associated with the variable transportation cost. Constraint (3.3b) represents both constraints (3.1g) and (3.1h), and the feasible region Ω is given by (3.2h)-(3.2x).

3.3 Accelerated Bilevel Decomposition Algorithm

As shown by Lara et al. (2018c), global optimization solvers do not perform well for mid to large instances of the single-period version of this problem. Thus, it is expected that with the added complexity of having multi-period decisions their performance will degrade even further. Lara et al. (2018c) propose a Bilevel Decomposition algorithm that consists of decomposing the problem into a master problem and a subproblem, for which ϵ-convergence can be proved. The master problem is based on a relaxation of the nonconvex MINLP, which yields an MILP that predicts the selection of facilities and their links to suppliers and customers, as well as a lower bound on the cost of the original problem. The subproblem corresponds to a nonconvex NLP of reduced dimensionality that results from fixing the binary variables in the MINLP problem, according to the binary variables predicted in the MILP master problem.

In this chapter, we propose an accelerated version of the algorithm proposed by Lara et al. (2018c) that keeps its rigor (i.e., its ϵ-convergence), but has some additional steps to improve its performance to allow the solution of large-scale multi-period instances of this problem within a reasonable amount of time. The additional steps consist of an attempt of reducing the optimization search space such that it is easier for the Bilevel Decomposition to find good bounds and the optimal solution. These steps consist of: i) possibly reducing the set of potential facilities by performing branch-and-bound on the facilities that were not selected; ii) potentially reducing the feasible two-dimensional space by performing a branch-and-bound on the partitions that did not have any facility being built on; iii) giving an initial feasible solution to the Master Problem based on the solution of the previous iteration.

The main steps in the Accelerated Bilevel Decomposition are shown in Figure 3.1.

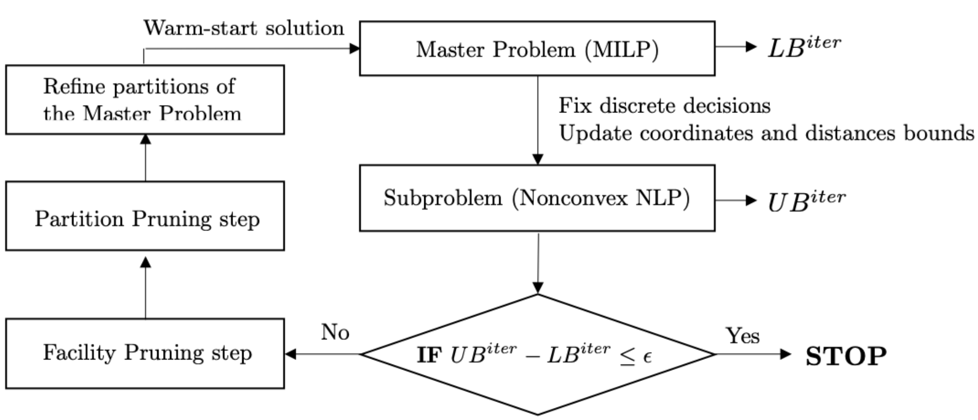

Figure 3.1: Accelerated Bilevel Decomposition concise representation

We start by explaining the basic steps of the original algorithm by Lara et al. (2018c) applied to the current MINLP formulation (3.2), and then cover the proposed additional steps to improve its performance.

3.3.1 Master Problem

The nonlinearity and nonconvexity of the formulation come from the fact that the location of the potential facilities is a decision variable. The master problem takes advantage of this property and partitions the space into uniform rectangular sub-regions. By having a grid to represent the feasible area, we can pre-compute the minimum distance between the fixed points (suppliers and customers) and use them as parameters in the model (Lara et al., 2018c). The minimum distances between the fixed points and the sub-regions p, $\widehat{D}_{i,p}$ and $\widehat{D}_{j,p}$, are computed as follows:

$$dx_{i,p} = \max\{|X_i - x_p| - \Delta x/2, 0\} \qquad \forall\, i \in \mathcal{I}, p \in \mathcal{P} \tag{3.4a}$$

$$dy_{i,p} = \max\{|Y_i - y_p| - \Delta y/2, 0\} \qquad \forall\, i \in \mathcal{I}, p \in \mathcal{P} \tag{3.4b}$$

$$dx_{j,p} = \max\{|X_j - x_p| - \Delta x/2, 0\} \qquad \forall\, j \in \mathcal{J}, p \in \mathcal{P} \tag{3.4c}$$

$$dy_{j,p} = \max\{|Y_j - y_p| - \Delta y/2, 0\} \qquad \forall\, j \in \mathcal{J}, p \in \mathcal{P} \tag{3.4d}$$

$$\widehat{D}_{i,p} = \max\{\sqrt{dx_{i,p}^2 + dy_{i,p}^2}, D^{min}\} \qquad \forall\, i \in \mathcal{I}, p \in \mathcal{P} \tag{3.4e}$$

$$\widehat{D}_{j,p} = \max\{\sqrt{dx_{j,p}^2 + dy_{j,p}^2}, D^{min}\} \qquad \forall\, j \in \mathcal{J}, p \in \mathcal{P}, \tag{3.4f}$$

where (x_p, y_p) are the coordinates of the mid-point of each sub-region p; Δx and Δy are the length of sub-region p in the x and y directions, respectively; D^{min} is the lower bound for the distances, not allowing the model to choose to build a facility k on top of a fixed point from a supplier or a customer (Lara et al., 2018c).

By using the minimum distance parameters, the MINLP formulation (3.2) can be reformulated as a mixed-integer linear programming (MILP) model (3.5), which yields a lower bound to the solution of the original models (3.1) and (3.2), as proved in Proposition 1 of Lara et al. (2018c).

$$\min\ \Phi = \sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t} \cdot \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}} \left(inv_{k,p,t} + op_{k,p,t} + \sum_{i\in\mathcal{I}} cost^{\mathrm{s}}_{i,k,p,t} + \sum_{j\in\mathcal{J}} cost^{\mathrm{c}}_{k,j,p,t} \right) \tag{3.5a}$$

$$\text{s.t}\quad inv_{k,p,t} = (FIC_{k,t} + VIC_{k,t}\cdot MC_k)\cdot b_{k,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5b}$$

$$op_{k,p,t} = VOC_{k,t}\cdot f_{k,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5c}$$

$$\begin{aligned} cost^{\mathrm{s}}_{i,k,p,t} &= CS_{i,t}\cdot ff^{\mathrm{s}}_{i,k,p,t} \\ &\quad + FTC^{\mathrm{s}}_{i,k}\cdot z^{\mathrm{s}}_{i,k,p,t} + VTC^{\mathrm{s}}_{i,k}\cdot \hat{D}^{\mathrm{s}}_{i,p}\cdot ff^{\mathrm{s}}_{i,k,p,t} \qquad \forall\, i\in\mathcal{I}, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \end{aligned} \tag{3.5d}$$

$$cost^{\mathrm{c}}_{k,j,p,t} = FTC^{\mathrm{c}}_{k,j}\cdot z^{\mathrm{c}}_{k,j,p,t} + VTC^{\mathrm{c}}_{k,j}\cdot \hat{D}^{\mathrm{c}}_{j,p}\cdot ff^{\mathrm{c}}_{k,j,p,t} \qquad \forall\, k\in\mathcal{K}, j\in\mathcal{J}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5e}$$

$$\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}} ff^{\mathrm{s}}_{i,k,p,t} \le AV_{i,t} \qquad \forall\, i\in\mathcal{I}, t\in\mathcal{T} \tag{3.5f}$$

$$\sum_{i\in\mathcal{I}} ff^{\mathrm{s}}_{i,k,p,t}\cdot CV_k = f_{k,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5g}$$

$$f_{k,p,t} = \sum_{j\in\mathcal{J}} ff^{\mathrm{c}}_{k,j,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5h}$$

$$\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}} ff^{\mathrm{c}}_{k,j,p,t} = DM_{j,t} \qquad \forall\, j\in\mathcal{J}, t\in\mathcal{T} \tag{3.5i}$$

$$w_{k,p,t} = \sum_{i\in\mathcal{I}} z^{\mathrm{s}}_{i,k,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5j}$$

$$w_{k,p,t} = \sum_{j\in\mathcal{J}} z^{\mathrm{c}}_{k,j,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5k}$$

$$w_{k,p,t} = w_{k,p,t-1} + b_{k,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5l}$$

$$\sum_{p\in\mathcal{P}} w_{k,p,t} \le 1 \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.5m}$$

$$\sum_{p\in\mathcal{P}} b_{k,p,t} \le 1 \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.5n}$$

$$\sum_{p\in\mathcal{P}} z^{\mathrm{s}}_{i,k,p,t} \le 1 \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.5o}$$

$$\sum_{p\in\mathcal{P}} z^{\mathrm{c}}_{k,j,p,t} \le 1 \qquad \forall\, k\in\mathcal{K}, t\in\mathcal{T} \tag{3.5p}$$

$$0 \le f_{k,p,t} \le MC_k\cdot w_{k,p,t} \qquad \forall\, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5q}$$

$$0 \le ff^{\mathrm{s}}_{i,k,p,t} \le \overline{FF^{\mathrm{s}}_{i,k,t}}\cdot z^{\mathrm{s}}_{i,k,p,t} \qquad \forall\, i\in\mathcal{I}, k\in\mathcal{K}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5r}$$

$$0 \le ff^{\mathrm{c}}_{k,j,p,t} \le \overline{FF^{\mathrm{c}}_{k,j}}\cdot z^{\mathrm{c}}_{k,j,p,t} \qquad \forall\, k\in\mathcal{K}, j\in\mathcal{J}, p\in\mathcal{P}, t\in\mathcal{T} \tag{3.5s}$$

$$\sum_{p} w_{k,p,t} \ge \sum_{p\in\mathcal{P}} w_{k+1,p,t} \qquad \forall\, k\in\mathcal{K}_n,\ n\in\mathcal{N}, t\in\mathcal{T} \tag{3.5t}$$

$$\sum_{p' \leq p} w_{k',p',t} \geq \sum_{p \in \mathcal{P}} w_{k,p,t} \qquad \forall\, k' < k,\ k, k' \in \mathcal{K}_n,\ n \in \mathcal{N}, t \in \mathcal{T} \tag{3.5u}$$

$$b_{k,p,t}, w_{k,p,t}, z^{\mathrm{s}}_{i,k,p,t}, z^{\mathrm{c}}_{k,j,p,t} \in \{0,1\} \qquad \forall\, i \in \mathcal{I}, k \in \mathcal{K}, j \in \mathcal{J}, p \in \mathcal{P}, t \in \mathcal{T}. \tag{3.5v}$$

Following the same notation as in (3.3), the MILP master problem (3.5) can be concisely represented by (3.6).

$$\Phi^{LB} = \quad \min \quad g(f, ff, z) + D^{\intercal} C ff \tag{3.6a}$$

$$\text{s.t.} \qquad f, ff, z, d, x, y \in \Omega' \tag{3.6b}$$

where D is the vector of minimum distance parameters, $\widehat{D}_{i,p}$ and $\widehat{D}_{j,p}$, and Ω' represents the feasible region described by constraints (3.5f)-(3.5v).

3.3.2 Subproblem

After solving the master problem (3.5), the subproblem consists of solving (3.2) for the fixed decisions of which facilities k to build and operate at each time period t, $\widehat{b}_{k,t}$ and $\widehat{w}_{k,t}$, respectively, and how to allocate their material supply $\widehat{z}^{\mathrm{s}}_{i,k,t}$ and products $\widehat{z}^{\mathrm{c}}_{k,j,t}$ as selected in the MILP (3.5).

Besides fixing the discrete decisions, we also update the bounds of the facilities coordinates such that their location (x_k, y_k) has to lie within the bounds of the sub-region p chosen in the Master Problem; i.e., for a p such that $\sum_{t \in \mathcal{T}} b_{k,p,t} = 1$ in the solution of Problem (3.5) we have that $\underline{X}'_{k,p} \leq x_k \leq \overline{X}'_{k,p}$ and $\underline{Y}'_{k,p} \leq y_k \leq \overline{Y}'_{k,p}$, where:

$$\underline{X}'_{k,p} = x_p - \Delta x/2 \qquad \forall\, k \in \mathcal{K} \tag{3.7a}$$

$$\overline{X}'_{k,p} = x_p + \Delta x/2 \qquad \forall\, k \in \mathcal{K} \tag{3.7b}$$

$$\underline{Y}'_{k,p} = y_p - \Delta y/2 \qquad \forall\, k \in \mathcal{K} \tag{3.7c}$$

$$\overline{Y}'_{k,p} = y_p + \Delta y/2 \qquad \forall\, k \in \mathcal{K}. \tag{3.7d}$$

This assumption greatly impacts tractability because the bounds for $d_{i,k}$ and $d_{k,j}$, which are part of the bilinear terms, become tighter, i.e., $\underline{D}'_{i,k,p} \leq d_{i,k} \leq \overline{D}'_{i,k,p}$ and $\underline{D}'_{k,j,p} \leq d_{k,j} \leq \overline{D}'_{k,j,p}$, where:

$$\underline{D}'_{i,k,p} = \widehat{D}_{i,p} \qquad \forall\, i \in \mathcal{I},\ k \in \mathcal{K} \tag{3.8a}$$

$$\overline{D}'_{i,k,p} = \widehat{D}_{i,p} + \sqrt{\Delta x^2 + \Delta y^2} \qquad \forall\, i \in \mathcal{I},\ k \in \mathcal{K} \tag{3.8b}$$

$$\underline{D}'_{k,j,p} = \widehat{D}_{j,p} \qquad \forall\, j \in \mathcal{J},\ k \in \mathcal{K} \tag{3.8c}$$

$$\overline{D}'_{k,j,p} = \widehat{D}_{j,p} + \sqrt{\Delta x^2 + \Delta y^2} \qquad \forall\, j \in \mathcal{J},\ k \in \mathcal{K}. \tag{3.8d}$$

Accordingly, the McCormick convex envelopes (McCormick, 1976) also become tighter, strengthening the lower bounds in the global optimization search of this NLP.

Following the same notation as in (3.3) and (3.6), the NLP subproblem can be concisely represented by (3.9).

$$\Phi^{UB} = \quad \min \quad g(f, ff, \hat{z}) + d^{\intercal} C ff \tag{3.9a}$$

$$\text{s.t.} \quad d_{l,k} \geq \sqrt{(X_l - x_k)^2 + (Y_l - y_k)^2} \qquad \forall l \in \mathcal{I} \cup \mathcal{J},\ k \in \mathcal{K} \tag{3.9b}$$

$$f, ff, d, x, y \in \Omega'' \tag{3.9c}$$

where $\hat{z}$ represents the discrete decisions obtained in the solution of the Master Problem 3.6 and fixed for this Subproblem, and Ω'' represents the feasible region Ω with the updated bounds for the distances and (x, y) coordinates, $d_{i,k}$, $d_{k,j}$, and (x_k, y_k), respectively.

The subproblem (3.9) is a reduced nonconvex NLP. Since it comprises the original problem (3.3) for a set of fixed discrete decisions and tighter bounds for the distances and (x,y) coordinates, it yields a feasible Upper Bound (UB) to the total cost, $\Phi^{UB} \geq \Phi$.

3.3.3 Facility Pruning

There are instances, especially the ones that favor centralized networks, in which having a large set of potential distributed facilities adds unnecessary burden to their solution. With this in mind, we propose an additional step to the original Bilevel Decomposition (Lara et al., 2018c) based on the branch-and-bound algorithm. After solving the Master Problem and the Subproblem, this step consists of solving the MILP (3.5) with the additional constraint:

$$\sum_{p \in \mathcal{P}} \sum_{t \in \mathcal{T}} b_{k',p,t} \geq 1 \tag{3.10}$$

for each facility k' that was not selected to be built by the Master Problem (3.5). Constraint (3.10) enforces facility k' to be built in one of the partition during the considered planning horizon. $\Phi^{LB,k'}$ is the optimal solution of the MILP (3.5) with constraint (3.10) for facility k'. Based on this additional constraint, we have the following Proposition 1.

Proposition 1. If the result of the MILP (3.5) plus the additional constraint (3.10), $\Phi^{LB,k'}$, is greater than the upper bound obtained by the NLP subproblem, it means that building this facility k' will never be optimal hence it can be excluded from the set of potential facilities.

Proof. From Proposition 1 of Lara et al. (2018c) we know that the optimal value of the MILP (3.5), Φ^{LB}, is an underestimator of the optimal value of MINLP (3.5), Φ. Hence, the optimal value of the MILP (3.5) with the additional constraint (3.10), $\Phi^{LB,k'}$, underestimates the optimal value of MINLP (3.5) with this additional requirement of forcing facility k' to be built within the planning horizon, $\Phi^{k'}$.

Moreover, from Theorem 1 of Lara et al. (2018c) we have that the optimal value of the NLP subproblem, Φ^{UB} is an incumbent (i.e. feasible solution) of the MINLP (3.5), such that $\Phi^{LB} \leq \Phi \leq \Phi^{UB}$. Therefore, if $\Phi^{LB,k'} > \Phi^{UB}$ and $\Phi^{LB,k'} \leq \Phi^{k'}$, then $\Phi^{k'} > \Phi^{UB}$ and, consequently, building this facility k' will never be optimal. Accordingly, facility k' can be pruned from the set of potential facilities. □

Since all facilities of the same type have exactly the same characteristics and data and the symmetry breaking constraint (3.2t) forces lower-index facilities of the same type to be build first, then if facility k' is pruned, it means that all facilities k'' such that $k'' > k'$ should also be pruned (i.e., excluded from the set of potential facilities).

This step can be computationally expensive; therefore we only perform it in the first iteration of the algorithm, and also set a maximum solution time for the solution of each $\Phi^{LB,k'}$.

3.3.4 Partition Pruning

The idea of the Partition Pruning step is very similar to the Facility Pruning. It consists of running a set of MILPs (3.5) with the additional constraint:

$$\sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} b_{k,p',t} \geq 1 \tag{3.11}$$

for each partition p' that did not have any facility k being built on by the Master Problem (3.5). This constraint enforces that at least one facility k is built on this partition p' during the planning horizon. $\Phi^{LB,p'}$ is the optimal solution of the MILP (3.5) with constraint (3.11) for partition p'.

Following the same idea as before, we can establish the following Proposition 2.

Proposition 2. If the result of the MILP (3.5) with the additional constraint (3.11), $\Phi^{LB,p'}$, is greater than the upper bound obtained by the NLP subproblem, it means that building on this partition p' will never be optimal and this partition and its further refinements can be excluded from the set of potential partitions.

Proof. This proof is very similar to the proof of Proposition 1. Following the same logic as before we have that the optimal value of the MILP (3.5) with the additional constraint (3.11), $\Phi^{LB,p'}$, underestimates the optimal value of MINLP (3.5) with this additional requirement of forcing at least one facility to be built on partition p', $\Phi^{p'}$. Therefore, knowing that $\Phi^{LB} \leq \Phi \leq \Phi^{UB}$, if $\Phi^{LB,k'} > \Phi^{UB}$ and $\Phi^{LB,k'} \leq \Phi^{k'}$, we can conclude that $\Phi^{k'} > \Phi^{UB}$ and, consequently, building a facility on partition p' will never be optimal. Accordingly, partition p' and its further refinements can be pruned from the set of potential facilities. $\square$

This step can also be computationally expensive; therefore we only perform it in the first two iterations of the algorithm, and also set a maximum solution time for the solution of each $\Phi^{LB,p'}$.

Additional to the Partition Pruning step, we automatically prune the partitions that have their minimum distance to the fixed points, $\hat{D}_{i,p}$ and $\hat{D}_{j,p}$, plus the diagonal size of the partition $\sqrt{\Delta x^2 + \Delta y^2}$ to be less than the allowed minimum distance D_{min}, which means that we prune the partitions in which the maximum distance between them and a fixed point is less than minimum distance allowed, which would violate the distance bound constraints in the original MINLP (3.2v) and (3.2w).

3.3.5 Warm-start MILP solutions

The solution of the MILP (3.5) is the main bottleneck to the solution of the Bilevel Decomposition algorithm because as the number of partitions increases, it greatly impacts the size of the model and, consequently, its solution time. In order to mitigate this issue, we warm-start the MILP solutions by providing to the solver a good feasible solution.

This initial feasible solution is directly obtained from the solution of the Master Problem and Subproblem in the previous iteration, not requiring to solve any additional MILP primal heuristic (Fischetti and Lodi, 2011). This feasible solution consists of building the facilities selected on the previous Master Problem (at the same time period as before), and choosing for their location the partition corresponding to the (x_k, y_k) coordinates given by the previous NLP Subproblem. In case the NLP Subproblem builds the facility on the boundary of the partition chosen by the MILP Master Problem, we select for the warm-start solution the adjacent partition that shares this boundary. This feasible solution is provided to the MILP solver (e.g. Gurobi and CPLEX) through the *initialize* option in Pyomo.

This step is not necessary, as we did not encounter any case in which the MILP solver could not find a feasible solution without the warm-start. Also, it does not reduce the computational time required by the MILP solver to solve the LP relaxation. However, it does provide a good incumbent solution that can help the convergence of the Branch-and-Bound algorithm, and expedite the overall convergence of the Accelerated Bilevel Decomposition, as can be seen in results in sections 3.3.8 and 3.4.

3.3.6 Accelerated Algorithm

As discussed earlier in this section, the Accelerated Bilevel Decomposition Algorithm consists of iteratively solving the MILP master problem and the NLP subproblem with additional steps to help convergence: Facility Pruning, Partition Pruning, and Warm-start of the Master Problem. The proposed algorithm is shown in Figure 3.2.

As proved by Theorem 1 in Lara et al. (2018c), the proposed Bilevel Decomposition algorithm in Figure 3.2 converges to the global optimum in a finite number of steps within an ϵ-tolerance.

3.3.7 Relation between space discretization and optimality tolerance

The lower bound of the algorithm is tightly related to how refined the discretization of space is in the current iteration, as the lower bound comes from the solution of the MILP (3.5) in which the distance variable is underestimated as the minimum distance between the fixed points and each partition on the grid. Therefore, Proposition 3 finds an upper bound to the dimensions of the partitions in the grid, Δ^*, such that if $\Delta x \leq \Delta^*$ and $\Delta y \leq \Delta^*$ the algorithm will converge in one iteration.

Proposition 3. By starting the Bilevel Decomposition algorithm with a specific $p_x^* \times p_y^*$ partitioning of the space such that $\Delta x \leq \Delta^*$ and $\Delta y \leq \Delta^*$, the algorithm converges within ϵ-tolerance in a single iteration.

Proof. This proposition is true if by starting the Bilevel Decomposition algorithm with a partitioning of the space such that $\Delta x \leq \Delta^*$ and $\Delta y \leq \Delta^*$, the Master Problem in $iter = 1$ yields an upper bound, Φ^{UB}, the Subproblem in $iter = 1$ yields a lower bound, Φ^{LB}, and both satisfy the optimality tolerance $\Phi^{UB} - \Phi^{LB} \leq \epsilon$. Thus, from (3.6) and (3.9) we have

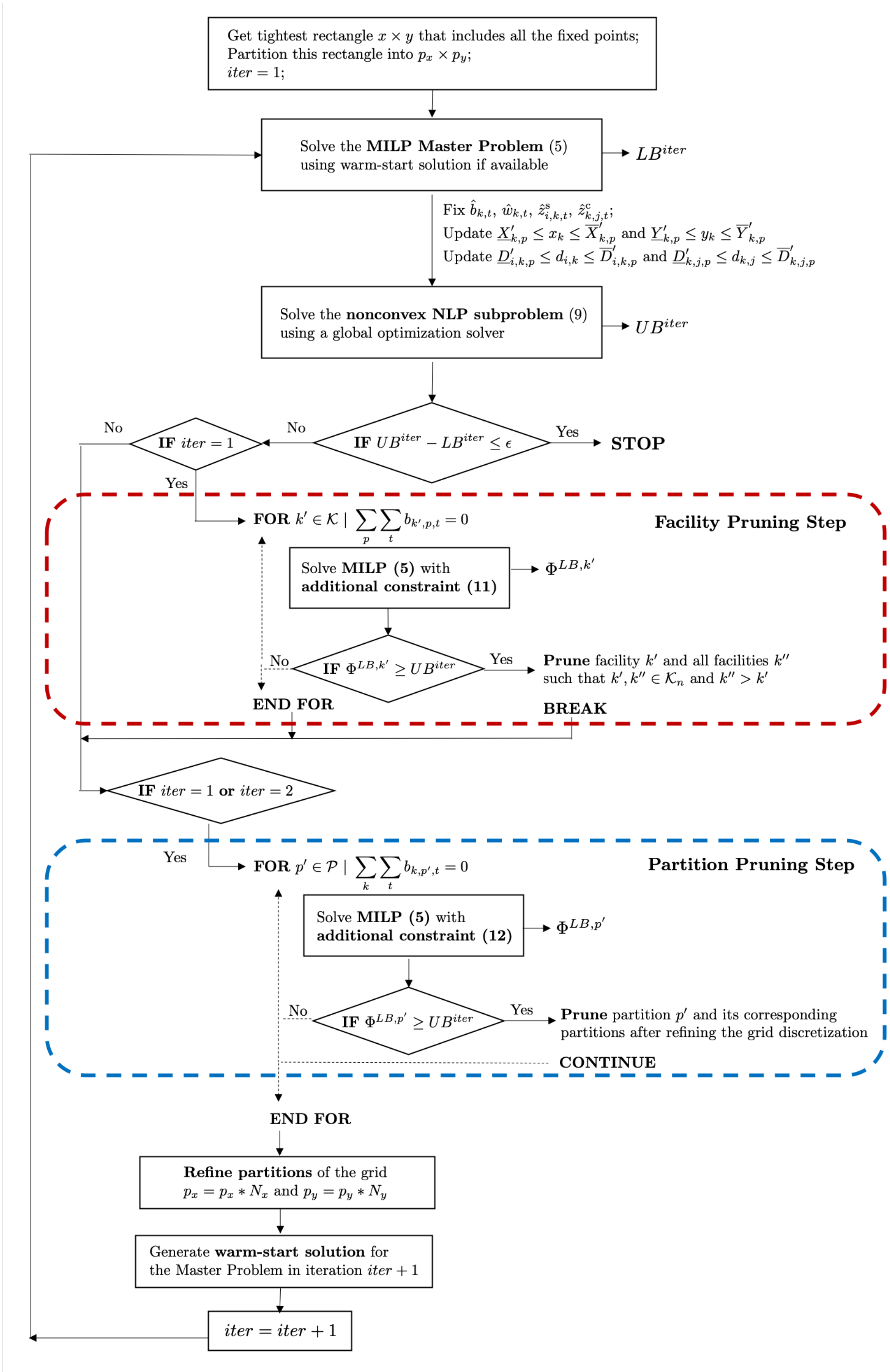

Figure 3.2: Accelerated Bilevel Decomposition Representation

that:

$$\Phi^{UB} - \Phi^{LB} \leq \epsilon \tag{3.12a}$$

$$g(f^*, ff^*, \hat{z}) + d^{*\intercal} C ff^* - g(\hat{f}, \widehat{ff}, \hat{z}) - D^\intercal C \widehat{ff} \leq \epsilon \tag{3.12b}$$

where the superscript * denotes the optimal solution of the variables in the NLP (3.9), and the accent $\hat{}$ denotes the optimal solution of the variables in the MILP (3.6).

From Proposition 1 by Lara et al. (2018c) we know that the difference between the MINLP (3.2) and the MILP (3.5) is the underestimation of transportation costs by the latter. Additionally, from Proposition 2 of the same paper, we have that for an infinite number of partitions the MILP (3.5) becomes an exact infinite dimensional representation of the MINLP (3.2) and both (3.2) and (3.5) have the same optimal solution $\Phi^* = \hat{\Phi}$.

Since the difference in the optimal value of the MINLP and its MILP underestimation is only due to the underestimation of the bilinear term, we can fix the optimal solution for the continuous variables f and ff to be the same between in MILP and the NLP, i.e., $f^* = \hat{f}$ and $ff^* = \widehat{ff}$, and this would give as a feasible solution $\Phi^{feas} \geq \Phi^{UB}$. Thus, if the optimality tolerance is satisfied by Φ^{feas}, it is also satisfied by Φ^{UB}. Therefore, for the sake of simplicity, we can omit the superscripts and write that

$$d^\intercal C ff - D^\intercal C ff \leq \epsilon \tag{3.12c}$$

$$\sum_{l \in \mathcal{I} \cup \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} C_{l,k,t} \cdot ff_{l,k,t} \cdot (d_{l,k} - D_{l,k}) \leq \epsilon \tag{3.12d}$$

Since $d_{l,m} - D_{l,m} \leq \sqrt{(\Delta x^*)^2 + (\Delta y^*)^2}$, if

$$\sum_{l \in \mathcal{I} \cup \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} C_{l,k,t} \cdot ff_{l,k,t} \cdot \sqrt{(\Delta x^*)^2 + (\Delta y^*)^2} \leq \epsilon$$

is satisfied, then (3.12d) will consequently be satisfied.

Therefore, we can write that

$$\sqrt{(\Delta x^*)^2 + (\Delta y^*)^2} \cdot \sum_{l \in \mathcal{I} \cup \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} C_{l,k,t} \cdot ff_{l,k,t} \ \leq \epsilon \tag{3.12e}$$

Considering an upper bound on the term multiplying the $\sqrt{(\Delta x^*)^2 + (\Delta y^*)^2}$, we denote it with the $^{\overline{\ \ }}$ accent. If the following condition is satisfied,

$$\sqrt{(\Delta x^*)^2 + (\Delta y^*)^2} \cdot \overline{\sum_{l \in \mathcal{I} \cup \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} C_{l,k,t} \cdot ff_{l,k,t}} \ \leq \epsilon \tag{3.12f}$$

then we can ensure that (3.12d) is satisfied.

Going back to the original (not concise) representation:

$$\sum_{l \in \mathcal{I} \cup \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} C_{l,k,t} \cdot ff_{l,k,t} = \sum_{t \in \mathcal{T}} \frac{1}{(1+R)^t} \sum_{k \in \mathcal{K}} \left(\sum_{i \in \mathcal{I}} VTC^{\mathrm{s}}_{i,k,t} \cdot ff^{\mathrm{s}}_{i,k,t} + \sum_{j \in \mathcal{J}} VTC^{\mathrm{c}}_{k,j,t} \cdot ff^{\mathrm{c}}_{k,j,t} \right) \tag{3.12g}$$

We can then take the maximum of the variable transportation costs over the facilities $k \in \mathcal{K}$ such that $\overline{VTC^{\mathrm{s}}}_{i,t} = \max_{k \in \mathcal{K}} VTC^{\mathrm{s}}_{i,k,t}$, and $\overline{VTC^{\mathrm{c}}}_{j,t} = \max_{k \in \mathcal{K}} VTC^{\mathrm{c}}_{k,j,t}$, and substitute these parameters back into (3.12g):

$$\sum_{l \in \mathcal{I} \cup \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} C_{l,k,t} \cdot ff_{l,k,t} \leq \sum_{t \in \mathcal{T}} \frac{1}{(1+R)^t} \left(\sum_{i \in \mathcal{I}} \overline{VTC^{\mathrm{s}}}_{i,t} \sum_{k \in \mathcal{K}} ff^{\mathrm{s}}_{i,k,t} + \sum_{j \in \mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} \sum_{k \in \mathcal{K}} ff^{\mathrm{c}}_{k,j,t} \right) \tag{3.12h}$$

Combining (3.12h) with constraint (3.2k) we can rewrite (3.12h) as follows

$$\sum_{l \in \mathcal{I} \cup \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} C_{l,k,t} \cdot ff_{l,k,t} \leq \sum_{t \in \mathcal{T}} \frac{1}{(1+R)^t} \left(\sum_{i \in \mathcal{I}} \overline{VTC^{\mathrm{s}}}_{i,t} \sum_{k \in \mathcal{K}} ff^{\mathrm{s}}_{i,k,t} + \sum_{j \in \mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} DM_{j,t} \right) \tag{3.12i}$$

and the only remaining variable is $ff^{\mathrm{s}}_{i,k,t}$.

Now, we can also take maximum value of the variable transportation cost over suppliers

i such that $\overline{\overline{VTC^{\mathrm{s}}}}_t = \max_{i\in\mathcal{I}} \overline{VTC^{\mathrm{s}}}_{i,t}$. Using this new parameter combined with constraints (3.2i) and (3.2k), and knowing that CV_k represents the conversion of facility k thus it is a fraction number between [0, 1] considering its minimum $\underline{CV} = \min_{k\in\mathcal{K}} CV_k$, we can rewrite (3.12i) as follows

$$\begin{aligned}
\sum_{l\in\mathcal{I}\cup\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}} C_{l,k,t}\cdot ff_{l,k,t} \quad &\leq \sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t}\left(\overline{\overline{VTC^{\mathrm{s}}}}_t \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{I}} ff^{\mathrm{s}}_{i,k,t} + \sum_{j\in\mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} DM_{j,t}\right) && (3.12\mathrm{j})\\
&\leq \sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t}\left(\overline{\overline{VTC^{\mathrm{s}}}}_t \sum_{k\in\mathcal{K}} \frac{f_{k,t}}{CV_k} + \sum_{j\in\mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} DM_{j,t}\right) && (3.12\mathrm{k})\\
&\leq \sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t}\left(\overline{\overline{VTC^{\mathrm{s}}}}_t \frac{\sum_{k\in\mathcal{K}} f_{k,t}}{\min_{k\in\mathcal{K}} CV_k} + \sum_{j\in\mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} DM_{j,t}\right) && (3.12\mathrm{l})\\
&\leq \sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t}\left(\frac{\overline{\overline{VTC^{\mathrm{s}}}}_t}{\underline{CV}} \sum_{j\in\mathcal{J}} DM_{j,t} + \sum_{j\in\mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} DM_{j,t}\right) && (3.12\mathrm{m})
\end{aligned}$$

With this result, we can go back to (3.12f) and rewrite it as:

$$\sqrt{(\Delta x^*)^2 + (\Delta y^*)^2} \leq \frac{\epsilon}{\sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t}\left(\frac{\overline{\overline{VTC^{\mathrm{s}}}}_t}{\underline{CV}} \sum_{j\in\mathcal{J}} DM_{j,t} + \sum_{j\in\mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} DM_{j,t}\right)} \tag{3.12n}$$

The last step can be applied since the costs and flows are positive, not affecting the sign of the inequality. Therefore, for $\Delta^* = \max(\Delta x^*,\ \Delta y^*)$, we can write

$$\Delta^* \leq \frac{\epsilon}{\sqrt{2}\sum_{t\in\mathcal{T}} \frac{1}{(1+R)^t}\left(\frac{\overline{\overline{VTC^{\mathrm{s}}}}_t}{\underline{CV}} \sum_{j\in\mathcal{J}} DM_{j,t} + \sum_{j\in\mathcal{J}} \overline{VTC^{\mathrm{c}}}_{j,t} DM_{j,t}\right)} \tag{3.12o}$$

Hence, if the user selects a partitioning of the space $p_x^* \times p_y^*$ such that $\Delta x \leq \Delta^*$ and $\Delta y \leq \Delta^*$ and Δ^*, and Δ^* is bounded by above as in (3.12o), then the Bilevel Decomposition algorithm converges in the first step. This means that the solution of the Master Problem in $iter = 1$ and the Subproblem in $iter = 1$ yield bounds that satisfy the optimality tolerance $\Phi^{UB} - \Phi^{LB} \leq \epsilon$. □

3.3.8 Illustrative Example

We illustrate how the algorithm works by solving a test-case using Network 2 from Lara et al. (2018c), with facility types 1 and 2 (centralized and distributed, respectively), 5 time-periods, 10% increase in demand by time period, and interest factor $R = 0.01$, and optimality tolerance of $\epsilon = 1\%$. We start $iter = 1$ with a $p_x = 2$ and $p_y = 2$ partition of the space, as shown in Fig. 3.3.

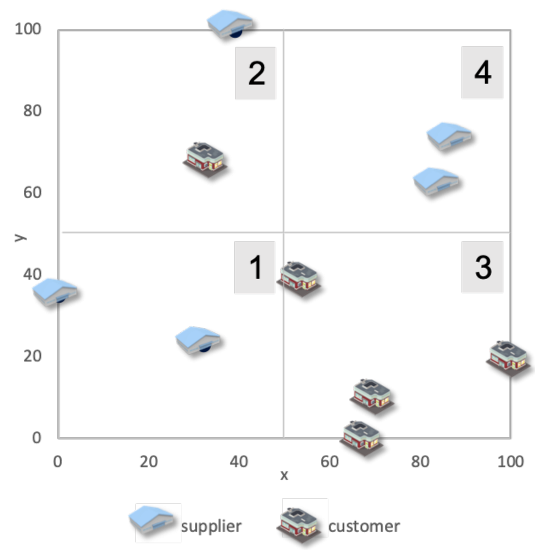

Figure 3.3: Illustrative problem: iteration 1 ($p_x = 2$, $p_y = 2$)

By solving the Master Problem (3.5) for this grid, we get a $LB^1 = 144,712$, and a solution that builds one centralized facility (type 1), $k = cf_1$, on partition $p = 2$ at time period $t = 1$ and keeps it operating throughout the planning horizon. We then solve Subproblem (3.9) for these fixed discrete decisions and obtain a solution that builds facility $k = cf_1$ on coordinate $(46.51, 70.76)$, yielding a feasible upper bound of $UB^1 = 149,236$, and an optimality gap of 3%. This gap is higher than the optimality tolerance, hence we proceed with the algorithm.

Since this is the first iteration, we perform the Facility Pruning step. We start by the second centralized facility $k = cf_2$ which was not built by the Master Problem. By solving the MILP (3.5) with the additional constraint (3.10), which enforces that $k = cf_2$ is built, we get $\Phi^{LB,cf_2} = 150,135$ which is higher than the current UB^1, thus we can prune $k = cf_2$ and know that the optimal solution does not have more than one centralized facility. We then continue to solve the Facility Pruning step for the distributed facilities. We start by solving the MILP with the additional constraint (3.10) for $k = df_1$ and get $\Phi^{LB,df_1} = 144,751$, which is lower than the current UB^1, thus cannot be pruned. We continue doing the same for $k = df_2$ and get $\Phi^{LB,df_2} = 146,267$, which is still lower than the UB^1. We then perform the same step for $k = df_3$ and get $\Phi^{LB,df_3} = 150,168$, which is higher than UB^1, thus we can prune $k = df_3$ and all the remaining distributed facilities, and know that the optimal solution does not have more than two distributed facilities.

The next step is to solve the Partition pruning. We solve the MILP (3.5) with the additional constraint (3.11) for $p = \{1, 3, 4\}$, and the results are shown in Table 3.1. Since none of the $\Phi^{LB,p}$ were higher than $UB^1 = 149,236$, we cannot prune any partition in this iteration. We proceed then to $iter = 2$, with $p_x = 4$ and $p_y = 4$ partition of the space, as represented in Fig. 3.4, keeping the updated set of potential facilities after pruning.

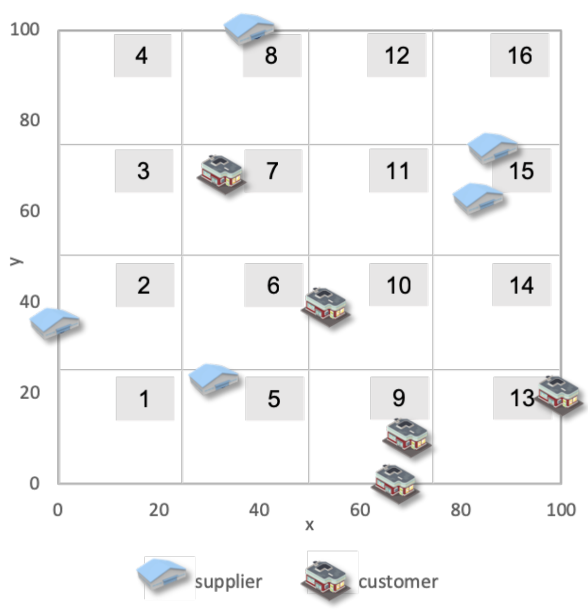

Figure 3.4: Illustrative problem: iteration 2 ($p_x = 4$, $p_y = 4$)

Based on the solution of the Master Problem and Subproblem for iteration 1, and the mapping between partitions in iterations 1 and 2, we warm-start the Master Problem MILP (3.5) with an initial feasible solution of building $k = cf_1$ on partition $p = 7$. The solution yields $LB^2 = 146,482$, and a solution that builds one centralized facility (type 1), $k = cf_1$, on partition $p = 11$ at time period $t = 1$ and keeps it operating throughout the planning horizon. We then solve Subproblem (3.9) for these fixed discrete decisions and get a solution that builds facility $k = cf_1$ on coordinate $(50.00, 69.97)$, yielding a feasible that is higher than the previous upper bound, so we keep $UB^2 = 149,236$. The optimality gap is now 2%, which is still higher than the optimality tolerance of 1%.

The following step is to solve the Partition pruning for the current grid. Since the Subproblem builds facility $k = cf_1$ on the boundary between partitions $p = 7$ and $p = 11$, we consider both of them as active and exclude them of the list of partitions to perform the Partition Pruning step. We solve the MILP (3.5) with the additional constraint (3.11) for $p = \{1, \ldots, 16\} \setminus \{7, 11\}$ and the respective results are shown in Table 3.1. Based on the results we can prune the current partitions $p = \{13, 14, 16\}$ and their further refinements.

Table 3.1: Partition Pruning step results (numbers in bold correspond to partitions that were pruned in the respective iteration)

	Iteration 1 $UB^1 = 149,236$	Iteration 2 $UB^2 = 149,236$
Partition p	$\Phi^{LB,p}$	$\Phi^{LB,p}$
1	146,845	148,146
2	-	147,964
3	144,828	148,146
4	144,828	148,892
5		148,093
6		147,452
7		-
8		146,781
9		149,364
10		147,828
11		-
12		146,994
13		**150,531**
14		**149,646**
15		148,688
16		**149,691**

We proceed to $iter = 3$, with $p_x = 8$ and $p_y = 8$ partition of the space, as represented in

Fig. 3.5. All partitions marked with a stripped pattern were pruned in the previous iteration.

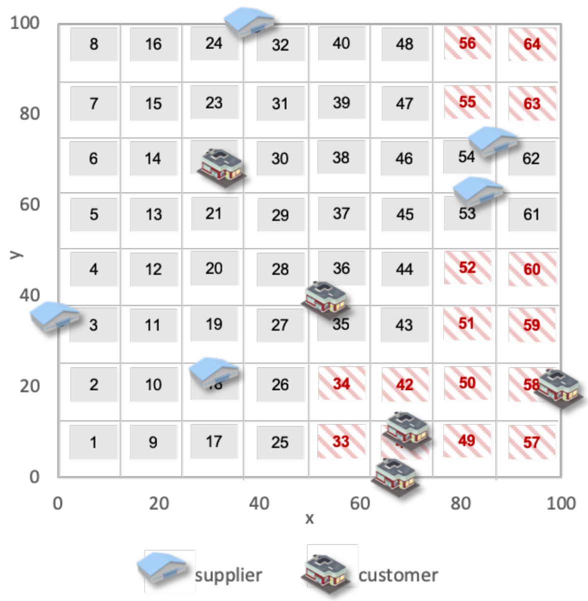

Figure 3.5: Illustrative problem: iteration 3 ($p_x = 8$, $p_y = 8$)

Using the solution of the Master Problem and Subproblem for iteration 2, and the mapping between partitions in iterations 2 and 3, we warm-start the Master Problem MILP (3.5) with an initial feasible solution of building $k = cf_1$ on partition $p = 30$. The solution yields $LB^3 = 147,805$, and a solution that builds one centralized facility (type 1), $k = cf_1$, on partition $p = 38$ at time period $t = 1$ and keeps it operating throughout the planning horizon. We then solve Subproblem (3.9) for these fixed discrete decisions and obtain a solution that builds facility $k = cf_1$ on coordinate $(50.00, 69.97)$, yielding a feasible that is higher than the previous upper bound, so we keep $UB^2 = 149,236$. The optimality gap is now 0.96%, which is lower than the optimality tolerance of 1%, therefore the algorithm has converged. The lower bound, upper bound and optimality gap at each iteration are reported in Table 3.2.

Table 3.2: Illustrative test problem results

iter	Lower Bound	Upper Bound	Gap
1	144,712	149,236	3%
2	146,482	149,236	2%
3	147,805	149,236	1%

As one can see, the lower bound gradually tightens up as the number of iterations *iter*, and consequently the number of partitions increase. The optimal network is shown in Figure 3.6. It takes 192 seconds to solve this instance on a macOS 2.3 GHz Intel Core i5, using Gurobi 8.0.1 to solve the MILPs (optimality tolerance of 0.01% for each MILP) and BARON 18.5.8 to solve the nonconvex NLP (time limit of 30 seconds per NLP). For the Facility Pruning and Partition Pruning steps, we limit the solution time of the MILPs to 10 seconds.

To evaluate the impact of each of the proposed steps, we solve the same instance using the Accelerated Bilevel decomposition: (i) without the Facility Pruning step, which takes 2770 seconds; (ii) without the Partition Pruning Step, which takes 211 seconds; and (iii) without the Warm-start step, which takes 196 seconds. This shows the proposed additional steps have an additive effect of the performance of the algorithm, and that the Facility pruning is the step with the greatest impact in the performance for this instance.

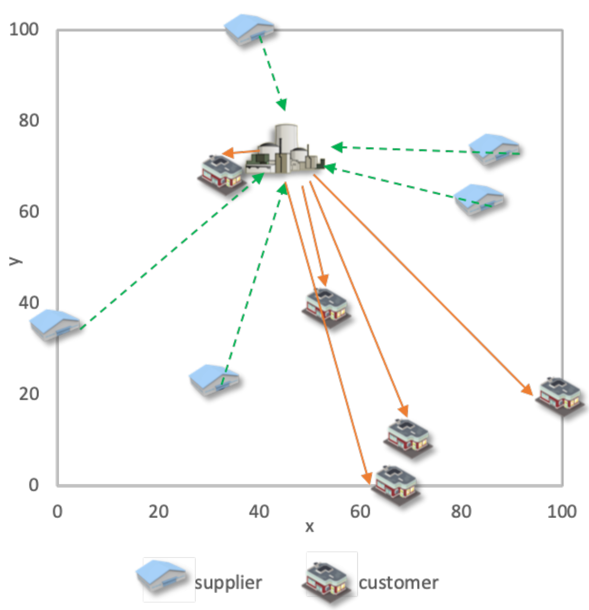

Figure 3.6: Illustrative problem optimal network

It takes 2,778 seconds to solve this same instance with the previous Bilevel Decomposition proposed by Lara et al. (2018c) using the same p_x, p_y, n_x and n_y. Additionally, BARON 18.5.8 takes 1,835 seconds to solve the original nonconvex MINLP (3.2) for this instance, while SCIP 5.0 and ANTIGONE 1.1 cannot solve it in 3,600 seconds (remaining optimality gaps of 2% and 4%, respectively).

By using Proposition 3, we get that if we start with $p_x, p_y \geq 20$ we have guaranteed convergence within 1% in the first iteration. This is considerably more refined than the $p_x, p_y = 8$ needed for the algorithm to converge, showing that even though Proposition 3 provides a valid bound, it is loose for this case, thus using it may add an unnecessary burden to the solution of the algorithm.

3.4 Computational results

In order to compare the performance of our proposed accelerated algorithm with the original algorithm and the currently available general purpose global optimization solvers, we 10 test cases from Lara et al. (2018c). The network varies in size as follows.

- Network 1: 2 suppliers × 2 consumers;
- Network 2: 5 suppliers × 5 consumers;
- Network 3: 10 suppliers × 10 consumers;
- Network 4: 20 suppliers × 20 consumers;
- Network 5: 40 suppliers × 40 consumers;

The 5 network structures are represented in Figures 2.6-2.10.

For each of the network options, we use as centralized facilities the Type 1 facilities from Lara et al. (2018c) (up to 2 large-scale facilities); and as distributed facilities, we first use Type 2 (up to 10 mid-scale facilities) and then Type 3 (up to 20 small-scale facilities). Therefore, for each of the network structures, the problem was solved for 12 and 22, respectively.

We assume that all instances are solved for 5 time periods, and the product demand and availability of raw material have a 10% increase per time-period.

Each test case is solved using the Accelerated Bilevel Decomposition (Fig 3.2), the original Bilevel Decomposition (Lara et al., 2018c), and by general purpose global optimization solvers, BARON, ANTIGONE and SCIP. We set the optimality tolerance to 2% and the maximum total CPU time to 1 hour. Regarding the algorithm, it is required that the Master Problem is solved to 0.5% optimality gap (and we use the lower bound of the MILP as the lower bound in the algorithm), and it is allowed a maximum CPU time of 30 seconds for the solution of each NLP Subproblem. We start the algorithm with a 2×2 partitioning of the space and at each iteration this partitioning is doubled, i.e $N_x, N_y = 2$.

Our computational tests were performed on a MacBook Pro laptop with a 2.3 GHz Intel Core i5, with 8GB of RAM, running on MacOS Mojave. We implemented the monolithic formulation and the global optimization algorithm in Python/Pyomo (Hart et al., 2017), solving the MILPs using Gurobi version 8.0.1 (Gurobi Optimization, 2018), the NLPs using BARON version 16.3.4 (Tawarmalani and Sahinidis, 2005), and the MINLPs using BARON version 18.5.8 (Tawarmalani and Sahinidis, 2005), ANTIGONE 1.1 (Misener and Floudas, 2014), and SCIP 5.0 (Gleixner et al., 2017). Source code reproducing our results is on Github (Lara, 2019).

The case-studies are named such that the first 2 letters represent the network (i.e., N1, N2, N3, N4, and N5, represent Network 1, 2, 3, 4, and 5, respectively), and the last 2 letters represent the facility types considered (i.e., T1T2 and T1T3 represent types 1 and 2, and types 1 and 3, respectively). The size of monolithic MINLP formulation (3.2) for each of the test cases is shown in Table 3.3.

Table 3.3: Monolithic MINLP formulation size

	Binary Variables	Continuous Variables	Constraints
N1-T1T2	360	393	1,265
N2-T1T2	720	825	2,087
N3-T1T2	1,320	1,545	3,457
N4-T1T2	2,520	2,985	6,197
N5-T1T2	4,920	5,865	11,677
N1-T1T3	660	703	2,925
N2-T1T3	1,320	1,495	4,407
N3-T1T3	2,420	2,815	6,877
N4-T1T3	4,620	5,455	11,817
N5-T1T3	9,020	10,735	21,697

The performance curves for the Accelerated Bilevel Decomposition, the original Bilevel Decomposition from Lara et al. (2018c) and each of the global solvers are shown in Figure 3.7.

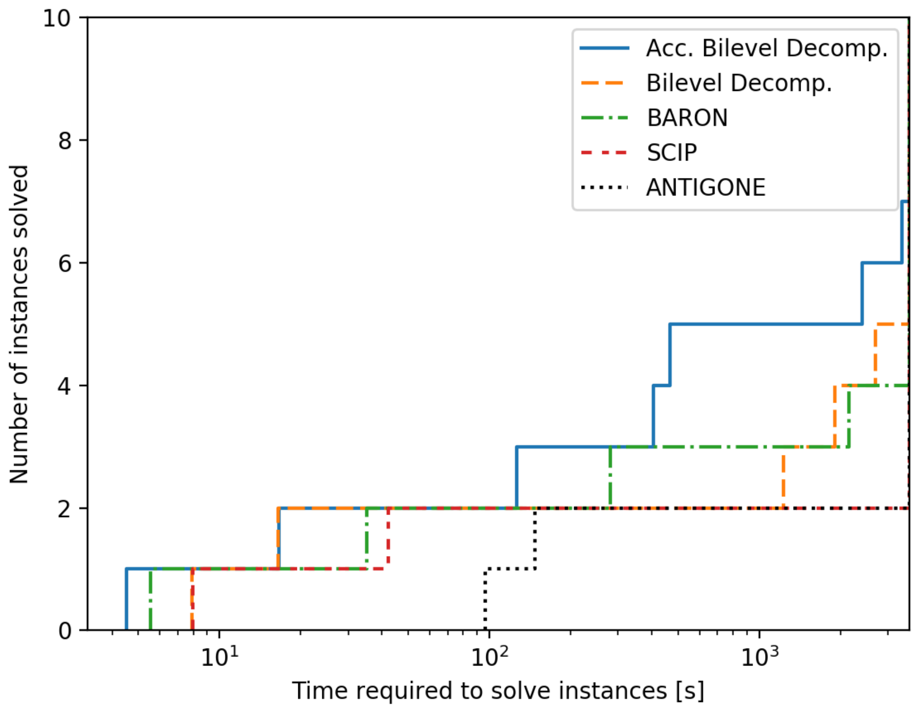

Figure 3.7: Performance curves comparing the Accelerated Bilevel Decomposition algorithm, with its original version and the commercial global optimization solvers.

The results show that the Accelerated Bilevel Decomposition algorithm was able to find the optimal solution within 2% optimality tolerance in 70% of the case studies, and performed better (i.e. found the optimal faster) than the other options in all of them. It can be noticed that there was a noticeable improvement in performance between the original Bilevel Decomposition and our Accelerated version of it, being able to solve 7 out of 10 instances instead of 5 out of 10. The global optimization solver that had the best performance for this problem and these instances was BARON. SCIP and ANTIGONE had a similar performance, only being able to solve 2 out of the 10 instances.

To evaluate the impact of each of the proposed steps, we solve these same 10 instances using the Accelerated Bilevel decomposition: (i) without the Facility Pruning step, (ii) without the Partition Pruning Step, and (iii) without the Warm-start step. The performance curves comparing these options against the proposed Accelerated Bilevel decomposition are shown in Figure 3.8.

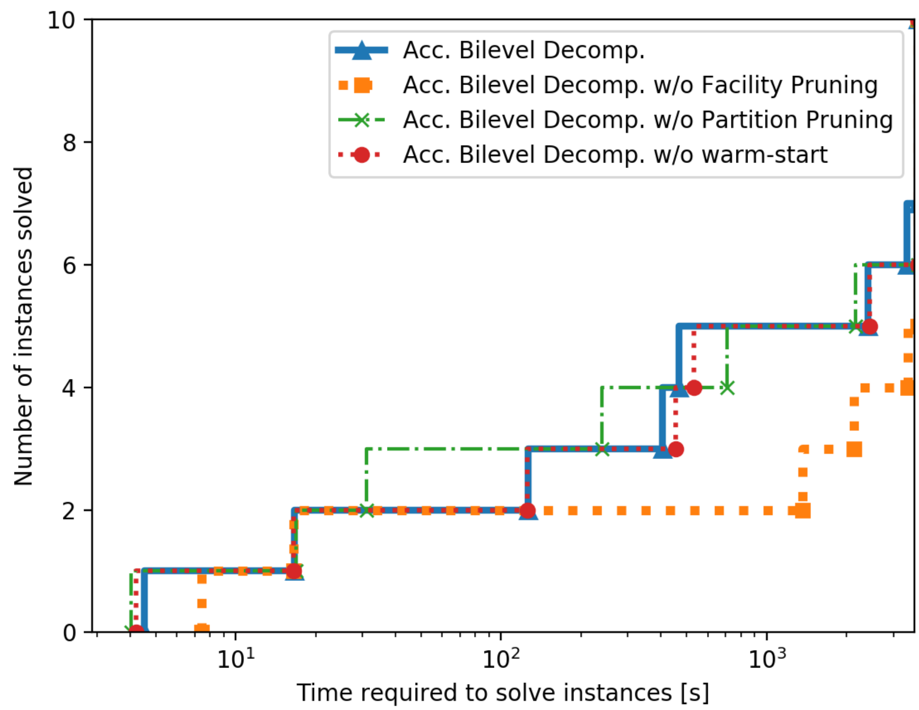

Figure 3.8: Performance curves comparing the Accelerated Bilevel Decomposition algorithm, with versions without Facility Step, Partition Pruning Step and Warm-start.

The results show that for the smaller instances the absence of each additional step did not have a great impact on performance. However, for larger instances each additional step was necessary to allow the solution of 7 instances. The algorithm without the Partition Pruning and without the Warm-Start could only solve 6 instances within 1 hour, and the algorithm without the Facility Pruning could only solve 5 instances within 1 hour, which shows that this is the step with the greatest impact in the performance. It is interesting to note that for smaller instances not having the Partition Pruning Step reduces the solution time, which makes sense since this can be a time consuming step that hurts the performance of easy instances.

3.5 Biomass supply chain case study

We present a bioethanol case study, adapted from the literature (Lara and Grossmann, 2016; Chen and Grossmann, 2019), to illustrate a real-world application for the proposed model and solution strategy. Given are 10 switchgrass suppliers and 10 ethanol markets with locations that are represented in Figure 3.9. There are 12 potential facilities to be built, of which 10 are distributed ($MC = 40.4$ MGal/year) and 2 are centralized facilities ($MC = 404$ MGal/year). All of the facilities have a conversion of $CV_k = 26\%$. Each market has a demand of 40 MGal of ethanol in the first year, with a 10% increase in demand each of the following years. Each supplier has 500 kilotonnes/year of switchgrass available, with a cost of \$30/ton, \$35/ton, \$33/ton, \$32/ton, \$37/ton, \$40/ton, \$34/ton, \$35/ton, \$31/ton and \$39/ton for suppliers 1 to 10, respectively. The fixed transportation costs ($FTC_{i,k,t}$, $FTC_{k,j,t}$) are \$10,000/year for all the possible links, and the variable transportation costs are \$2/ton-mile for the switchgrass ($VTC_{i,k,t}$) and \$0.40E-3/gal-mile for the ethanol ($VTC_{k,j,t}$). We solve this problem for a 5-year planning horizon.

The resulting model has 3,457 constraints, 1,545 continuous variables, and 1,320 binary variables. Starting with $p_x, p_y = 5$ and $N_x, N_y = 2$ it takes 3 iterations and 6 hours to solve it with the Accelerated Bilevel Decomposition within 2% optimality gap, with an optimal value of \$2.178 billion. We attempted to solve this same instance with BARON (the commercial global solver that has the best performance in the computational experiments in Section 3.4), but it only achieved 68% optimality gap when it reached the maximum solution time of 10 hours, highlighting again the need for a specialized algorithm such as the proposed Accelerated Bilevel Decomposition to be able to solve real-world applications of this problem.

The optimal network for the biomass supply chain problem is shown in Figure 3.10 (without the allocation links since it changes according to the time period). All the 10 distributed facilities were built in year 1, and one centralized facility was built in year 2. It is interesting

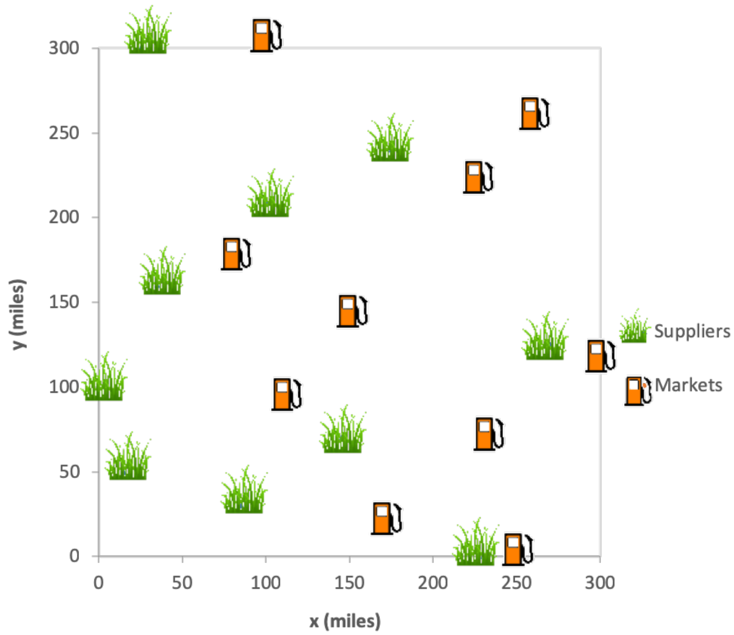

Figure 3.9: Network structure of the biomass supply chain (Lara and Grossmann, 2016)

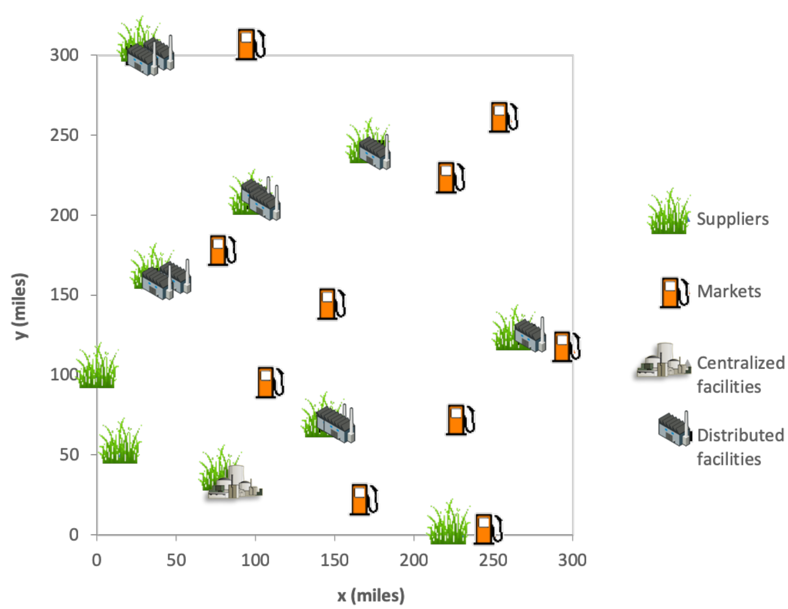

Figure 3.10: Optimal network for the biomass supply chain

to notice that in some cases the optimization decides to build 2 distributed modular plants right next to each other instead of replacing them with a larger-scale centralized plant.

3.6 Conclusions

This chapter has highlighted the need for a general model to optimize the design and planning of Distributed and/or Centralized manufacturing networks. We propose a GDP formulation to solve this problem, which belongs to the class of Capacitated Multi-facility Weber Problem.

We show that with the added complexity of having multi-period decisions the original Bilevel Decomposition proposed by Lara et al. (2018c) and the available global optimization solvers (BARON, ANTIGONE and SCIP) do not perform well, taking a long time to find feasible solutions and an acceptable optimality gap. Therefore, we propose an accelerated version of the Bilevel Decomposition with additional steps: Facility Pruning, Partition Pruning and Warm-start of the Master Problem. The additional steps do not compromise the rigorousness of the algorithm, which still has ϵ-convergence as proven in Lara et al. (2018c). We discuss theoretical properties of the algorithm and find an upper bound to the space discretization such that if the space is partitioned in any finer grid, the algorithm is guaranteed to converge in a single iteration.

Additionally, we perform computational experiments for the multi-period version of the random instances from Lara et al. (2018c), and show that the proposed Accelerated Bilevel Decomposition outperforms the original Bilevel Decomposition proposed by Lara et al. (2018c) and the available global optimization solvers (BARON, ANTIGONE and SCIP) in all the instances. Finally, we illustrate the applicability of the model and algorithm by solving a biomass supply chain problem from the literature.

Part II

Electric Power Infrastructure Planning

Chapter 4

Impact of Model Resolution on Scenario Outcomes for Electric Power Generation Expansion

In his chapter, we perform a systematic comparison of two alternative Generation Expansion Planning (GEP) frameworks to quantify how the choice of temporal representation and operational detail in a GEP model impacts the resulting capacity mix projections as well as operational metrics such as unmet demand, curtailment, and renewable energy penetration. Our analysis is based on a power system that approximately represents the U.S. Electric Reliability Council of Texas (ERCOT) grid in 2015. Since we focus on highlighting and understanding the reasons for the relative differences in the outputs of alternate GEP frameworks, the results presented here should not be interpreted as a detailed analysis of the ERCOT system along the lines of other efforts in the literature (e.g. Newell et al. (2014)). Indeed, transmission constraints both within and to and from ERCOT to other regions are not included (see Section 4.1). Grid stability at every instant of time is not verified. Energy storage is not considered. Revenue sufficiency constraints are not imposed for any generators. Therefore, all scenarios evaluated here are meant to exclusively highlight differences in GEP outputs and should not be interpreted as ERCOT grid projections.

The rest of the chapter is organized as follows. Section 4.1 provides a brief description of the two GEP models and the Production Cost Simulation (PCS) model, with comprehensive algebraic mathematical modeling details in the Appendix A. This is followed by a discussion of the main data inputs and assumptions in Section 4.2, including the methodology for generating representative time slices and days for the time-slice (TS-GEP) and chronological (C-GEP) models respectively. Section 4.3 presents the results of the inter-model comparison, along with the discussion of the impact of increasing the number of representative days in the C-GEP. Section 4.4 summarizes the key conclusions and policy implications.

4.1 Methods

Figure 4.1 summarizes the methodology used to investigate the impact of temporal resolution and operational detail in a GEP. Several works (e.g. Brinkman (2015) and Lew et al. (2013)) have focused on evaluating scenario outcomes of high renewables penetration using a single GEP (thus, these would follow the left or the right side of Figure 4.1 in which a single GEP is evaluated). More recently, a few analyses (Cole et al., 2017; Bistline et al., 2017) have attempted to compare the capacity projections across GEP models for a few scenarios, but have stopped short of comparing operational metrics associated with the projected capacity in a detailed simulation of annual grid operations. The analysis presented here relies on a common input data set comprising generator performance attributes and costs, load and renewable energy generation across multiple historical years and other parameters (see Section 5.5 and Appendix A). Based on this data set, we define the input parameters for each GEP, in particular load and renewable energy generation for the specific temporal resolution in each model. Subsequently, we evaluate the TS-GEP and C-GEP for a range of hypothetical renewable energy scenarios to estimate the generation capacity in future years. We go beyond prior GEP assessments by exploring the operational outcomes of the projected capacity mix of each GEP for multiple realizations of annual load and renewable energy generation profiles, using the PCS model. For each renewable penetration scenario, the PCS model simulates the annual hourly dispatch to meet demand and reports various metrics of interest, such as unmet demand, share of renewable energy generation, curtailment and so on. Together, the differences in operational metrics and generation capacity outputs across the two GEP models form the basis for our conclusions on the impact of temporal resolution in a GEP framework.

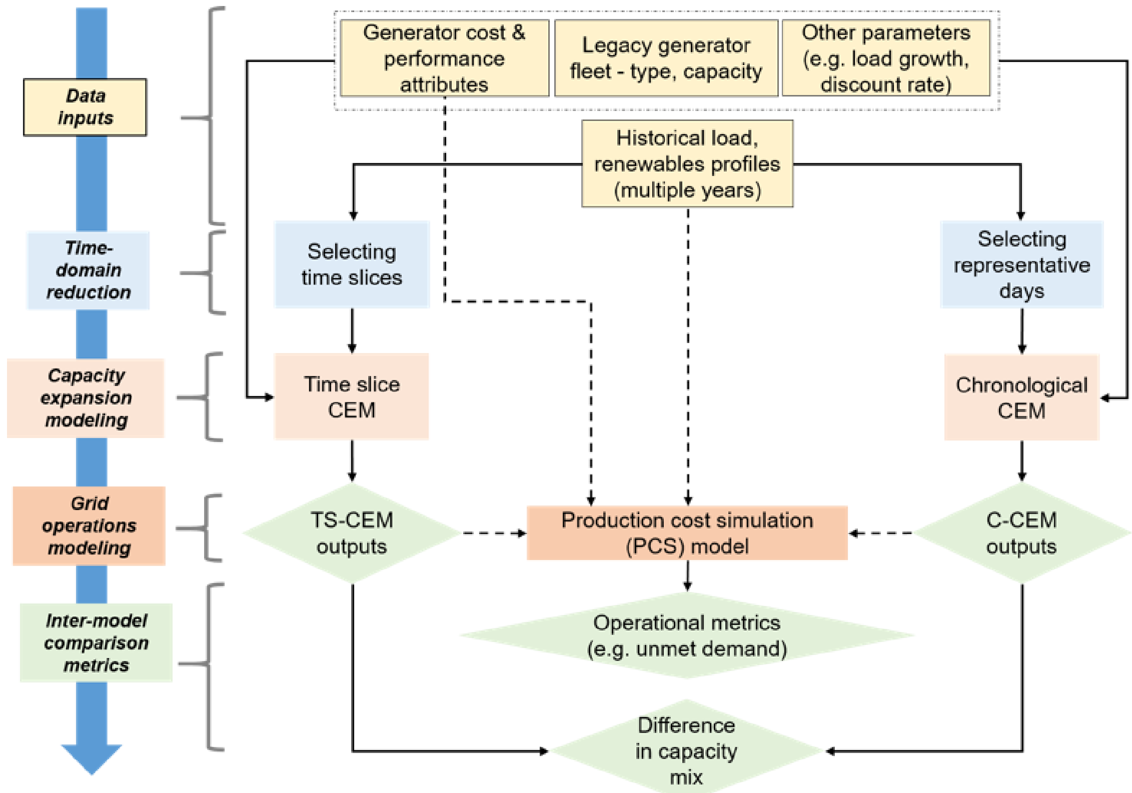

Figure 4.1: Methodology to systematically compare alternate temporal resolution in power system capacity expansion models across different scenarios. Diamonds refer to model outputs and dotted lines correspond to inputs to the production cost simulation (PCS) model. GEP = Capacity expansion model; TS = Time slice; C = Chronological.

4.1.1 Summary of Chronological and Time Slice Capacity Expansion Models

The two least-cost power generation GEP models developed in this study, C-GEP and TS-GEP, are deterministic inter-temporal optimization models that take the vantage point of a centralized planner seeking to determine cost-optimal expansion decisions over a planning horizon of several decades. Regarding their similarities, both models minimize total cost (discounted to present value), which includes installation (CAPEX) costs for new capacity being built, costs to extend the lifetime of installed capacity, operating costs (fixed and variable), fuel costs, generator start-up costs (C-GEP only), and unserved load. The objective function also includes cost savings attributed to the December 2016 implementation of the investment and production tax credits for wind and solar photovoltaic (PV) generation in the U.S. context (see data inputs in Table A.13) (DOE, 2016a,b). The models represent solar and wind capacity expansion decisions as "continuous" decisions meaning that a fractional wind generator can be built. Importantly, both models represent the existing fleet of coal, NG and nuclear as well as wind and solar PV generators in ERCOT by clustering the entire fleet into seven different generator types. As done in EPA (2013), both models also allow for aging capacity to be retired or extended, whereby the extension option incurs a one-time cost of extension and returns to operation with the same operational parameters. For each generation technology, both models include annual capacity installation limits (EPA, 2013) that implicitly account for supply chain constraints associated with emerging technology deployment. The models use as input the same forecasted load growth, the same suite of generation technologies to meet this growth, and the same associated cost assumptions to model grid evolution in 3-year time increments from 2015 to 2045 (see Appendix A.2). It is precisely the dissimilarities described below that will help elucidate why different expansion decisions are made by the C-GEP and TS-GEP in certain scenarios. Figure 4.2 illustrates the differences in temporal representation of grid operations between the C-GEP and TS-GEP. Temporally, the C-GEP represents annual load as well as wind and solar generation using 12 representative days at an hourly time resolution, whereas the TS-GEP represents annual load as well as wind and solar generation, with 16 time slices representing different times of day and seasons. In other words, the TS-GEP averages load and renewable energy capacity factor data (see Figure 4.3) in each of the four seasons (spring, summer, fall, winter) into time slices representing morning (7 am - 2 pm), afternoon (2-6 pm), evening (6-11 pm), and night (11 pm - 7 am). With the exception of minimum turndown constraints for coal and nuclear generators, discussed in Section 4.1.3, the TS-GEP does not link two consecutive time slices with respect to operational constraints.

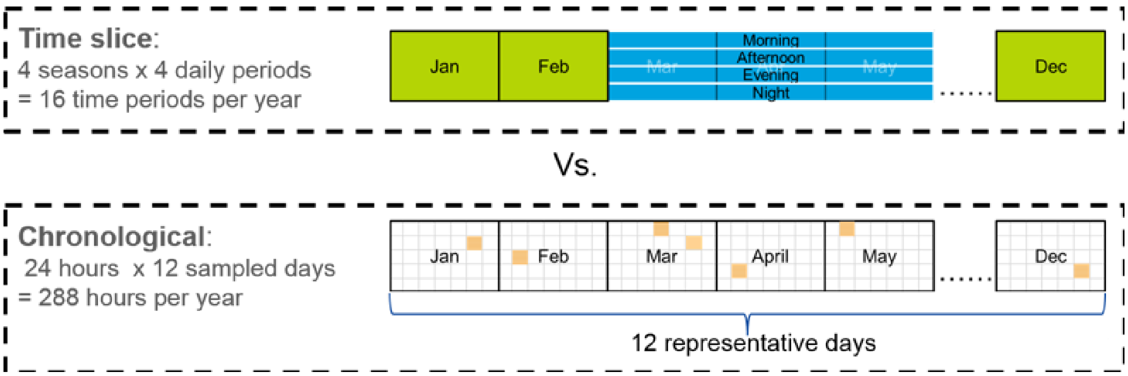

Figure 4.2: Differences in temporal resolution of the alternate capacity expansion models studied here: TS-GEP (top) and C-GEP (bottom).

In contrast, the C-GEP, as its name suggests, "sees" chronology, and therefore events that occur in a given hour are related to events that occur in the preceding and subsequent hours. The C-GEP is a deterministic mixed-integer linear program (MILP), while the TS-GEP is a deterministic linear program; both have perfect foresight. The complete mathematical formulation of the C-GEP and TS-GEP and data inputs are available in the supplementary information (Appendix A).

Operationally, the C-GEP considers important details associated with thermal generators including: unit commitment decisions (i.e. on/off commitments of generators to meet load), hourly ramping constraints, spinning reserves, quick-start reserves, and start-up costs. In contrast, the TS-GEP omits these details, although spinning reserves are partially taken into account. Lastly, because the C-GEP includes unit commitment decisions, thermal generation expansion decisions are modeled as integer decisions, unlike the TS-GEP which allows for a fractional number of thermal generators to be built. Together, the different temporal resolution and operating constraints of the two models lead to different generator dispatch profiles to meet load in each time period, which ultimately impacts the capacity investment decisions.

4.1.2 Production cost simulation (PCS) model

As is typically done in long-term expansion studies (see, e.g. Deane et al. (2012)), we use a PCS model, essentially a grid operations model, to independently assess annual operations (e.g. generation mix, unmet reserves, unmet load, and curtailment) given the installation decisions prescribed by the two GEP models. The PCS model simulates one year of grid operations at an hourly resolution. It takes as input a given installation of generators for a particular year (as determined by solving the GEP models) and solves a simultaneous unit commitment and economic dispatch problem over all 8760 hours in that year. Since the GEP models must approximate load and other features in order to be computationally tractable, the PCS model allows for extreme situations such as peak load, minimum renewable energy output, which may not be considered by the GEP models, to be evaluated. Further, for each projected generation capacity mix, we evaluate annual grid operations via the PCS model for seven different realizations of load, wind and solar PV capacity factor profiles, based on historical load and renewable generation data available from ERCOT. This approach allows for quantifying the robustness of capacity mix projected by the two GEP models to prevailing intra and inter-annual variability in load, wind and solar generation (see Section 4.3.2 and 4.3.3).

The PCS model is a deterministic MILP with perfect foresight. It is quite similar to the C-GEP, except that it does not include any installation related decisions and it models a single year of operations with hourly granularity, as opposed to several representative days. As with the C-GEP, generators are clustered by generator type. Generators in a cluster have the same attributes. This allows on/off decisions for thermal generators to be modeled using integer decision variables as opposed to binary decision variables, which would be needed if each individual thermal generator were modeled. The PCS is solved using a rolling horizon heuristic with a one-day overlap in which unit commitment and economic dispatch decisions are made (optimally decided by the mixed-integer programming solver) for 21 consecutive days (504 hours), after which the first 20 days are implemented. The model then "rolls forward" to considering the next 21 days. Thus, days 1-21 are optimized, followed by days

21-41, and so forth. The one-day overlap allows the model to "correct" decisions made in the last 24 hours that may have been too myopic due to the lack of additional foresight. Each sub-problem (21-day horizon) is solved with a 30-second time limit and a relative optimality gap tolerance of 0.01%.

A key difference between the two GEP models and the PCS is the presence of unmet load and unmet reserves. Because the GEP models "see" only a coarse representation of load and renewables profiles, they prescribe installation decisions that avoid instances of unmet load based on the limited operational data present in the model. However, the PCS may encounter more extreme variability in net load and thus may need to shed some reserves, or worse, load. The PCS penalizes unmet load at \$9000/MWh and approximates the penalty associated with unmet reserves using a step function, based on an approximation of the current operating reserve demand curve being used by ERCOT in clearing the market (Potomac Economics, 2016). In particular, for the case when the total operating reserves of 7.5% of load are desired for every hour, the first 1.5% of unmet operating reserves has a cost of \$100/MWh, the next 2% has a cost of \$300/MWh, the next 2% has a cost of \$3000/MWh, and the last 2% has a cost of \$9000/MWh. As a consequence, if an extreme net load event occurs, the PCS first sheds reserves according to this step function before ultimately shedding load.

The PCS model does not consider a real-time market. In contrast, commercially-available grid operations models like PLEXOS or GE MAPS that are designed to closely mimic power system operations, solve for the cost-optimal dispatch for one day ahead and subsequently re-adjust the dispatch (without turning on or off generators) based on new information (e.g. improved weather, load forecast) available in real-time conditions (e.g. 4 hours prior to dispatch) (Anderson et al., 2016). Another limitation of the PCS model is that it does not represent the minimum on and off times of thermal generators, which could overestimate their ability to respond to changing load and renewable energy generation and consequently underestimate the extent of curtailment reported for each scenario. Despite these limitations, the PCS model provides a consistent basis for comparing the operational differences in the capacity mix projected by the two GEP models.

4.1.3 Key model constraints of GEP models and PCS

A summary of the key constraints in the GEP models and the PCS model that highlight their similarities and differences are presented in Table 4.1. The complete mathematical formulation of the two GEP models are presented in Appendix A. A "time period" refers to an hour of a representative day in the C-GEP, to a time slice, i.e. a season-time block pair (e.g. summer-afternoon) in the TS-GEP, and an hour of the 8760 hours simulated in the PCS model.

Table 4.1: Key constraints present in the capacity expansion and production cost simulation models

Constraint type	Time Slice (TS-GEP)	Chronological (C-GEP)	Production Cost Simulation (PCS)
Load balance	X	X	X
Generator capacity balance	X	X	
Retirement or life extension decisions	X	X	
Annual installation limits	X	X	
Renewable Portfolio Standards	X	X	
Capacity planning reserve requirements	X	X	
System spinning and total operating reserve requirements	X	X	
Power output from generators upper and lower bounds		X	X
On/off commitment decisions		X	X
Generator ramping limits		X	X
Quick start and spinning reserves provided by thermal generation		X	X
Minimum turndown constraints to limit differences in seasonal max and min outputs	X		

The types of constraints governing investment and capacity decisions are:

- Capacity balance constraints. The capacity (MW) of each energy type in year y must equal the capacity in the previous year (y-1) plus the capacity that came on-line in year y minus the capacity that was retired in year y.
- Annual installation limits. There is an upper bound on the amount of capacity that can be built in each year for each technology type (see Table A.12).

- Retirement/extension constraints. Generators must not exceed their lifetime. They must be extended (with a cost penalty, see Section A.2 and Table A.10) or retired before or once they have reached their lifetime. Retired capacity immediately exits the fleet, i.e. is not available for generation. Extended capacity is kept for the remainder of the planning horizon and preserves all of the defining characteristics (e.g. heat rate), except its age is reset to zero.
- Minimum capacity reserve constraints. Minimum capacity reserve requirements are included to ensure that there is sufficient capacity to meet forecasted peak load in every year plus some margin of error (see Table 4.3). Capacity values are used to determine the contribution of each generator type to meet this constraint. Thermal generators are assumed to have a capacity value of one, while renewable energy generator types have a capacity less than one (see Section A.2, Table A.10 and Table A.11).
- Renewable Energy constraints. In the years when this constraint is implemented, total generation from renewables used to serve load must be at least a pre-defined percentage of the total system load in that year and beyond. This constraint is implemented to consider renewable energy scenarios (40% to 70%) where applicable.

The types of constraints that govern optimal unit commitment and economic dispatch include:

- Load constraints. In each time period, the total power (dispatched from thermal generators and generated from renewable energy) plus unserved load must equal total system load plus curtailment.
- Ramping constraints. These constraints, typically apply to thermal generators, and limit the increase/decrease in generation from one period to the next, i.e. rate of change in generation. They are particularly important when modeling generator operations at an hourly time scale when substantial load or renewable energy output fluctuations in successive periods are present. These are explicitly considered in the C-GEP and PCS model.
- Individual generator constraints. When operating, each generator produces power

between a minimum and maximum generation level. Because generator types are modeled, as opposed to individual generators, it is assumed that all units of a given type provide the same level of power and reserves capacity.

- Quickstart reserves. Thermal generators that are off during a given hour may contribute a fraction of their capacity to quickstart reserves. Total quickstart and spinning reserves must provide the necessary operating reserves (typically a % of load, say 7.5%) in every hour.
- Minimum turndown constraints. Due to its lack of high temporal resolution and chronology, the TS-GEP could choose to cycle, i.e. turn on and off, thermal plants at a higher frequency than permitted in practice. Minimum turndown constraints prevent this undesirable behavior by enforcing the following requirement: in each season, a thermal generator's power output in a time slice must be at least a pre-defined fraction of the maximum power output in that season. These are the only constraints in the TS-GEP that link operations in different time slices.

4.1.4 Modeling limitations

We focus our study on improving two aspects – operation detail and temporal resolution – of GEP models to evaluate scenarios of increasing renewable energy penetration. However, there are other areas of improvement that have deliberately not been addressed here. We did not consider transmission in either GEP to restrict the analysis to the impact of the temporal resolution differences (as opposed to spatial resolution differences) between the models. Moreover, transmission constraints were omitted to ensure the models, particularly the C-GEP, can be solved to optimality while looking ahead for the 30 year planning horizon with available solution algorithms. Transmission constraints have been ignored by other studies (Deetjen et al., 2016) evaluating grid expansion in the ERCOT context, with the justification that there is ample transmission in the region as a result of the recently completed transmission upgrades connecting the Panhandle area to the rest of ERCOT. Additionally, the two GEP models and the PCS model do not consider uncertainty (beyond typical reserve constraints), outages (planned or forced), storage, demand response, distributed generation, nor heat rate deterioration due to partial operation of thermal generators. Although these deficiencies may seem many, they apply to numerous leading models used in practice, e.g. IPM, NEMS, and ReEDs.

4.2 Case study and data inputs

4.2.1 Summary of data inputs

The power system we analyze as a case study approximately represents the ERCOT grid. We approximately model the existing ERCOT generator fleet by clustering individual generators into seven generator types as per the capacities as of May 2015 (ERCOT, 2015a): coal, nuclear, NGST, NGCC, Natural gas combustion turbine (NGCT), solar PV (single axis tracking) and wind. For simplicity, we do not include the relatively small amount of biomass and hydro capacity present within ERCOT. Electricity imports or exports to and from the region are ignored in this analysis, given their relatively small share of system demand for ERCOT (Mann et al., 2017). Within each cluster, there are an integer number of generators whose operating parameters are assumed to be the same. Table 4.2 summarizes the total installed capacity for each cluster type and the number of plants within that cluster. Each of the seven generator clusters are associated with an age distribution, based on the age of the individual generators that belong to that cluster as of 2015. This information, coupled with the economic lifetime of each generator type, is used in the TS-GEP and C-GEP to make retirement or extension decisions on the portion of the capacity that is scheduled to retire in each year of the planning period.

For new capacity additions, both GEP models are allowed to choose from nine different generator clusters: coal with and without carbon capture and sequestration (CCS), NGCC with and without CCS, NGCT, nuclear, new solar PV (single axis tracking), new solar concentrated solar thermal power (CSP) generation and new wind. The projected capital, variable operating costs and fixed costs over time for each generator type are retrieved from

Table 4.2: Summary of existing (2015) generator capacity used in the TS-GEP and C-GEP

System parameters	Installed capacity (MW)	Number of plants
Total	96,996	
Nuclear	5,164	4
Coal	17,397	27
NGCC	39,527	55
NGCT	7,481[1]	48
NGST	6,219	10
Solar PV	663	17
Wind	20,545	153

the NREL 2016 technology baseline database (NREL, 2016). Fuel costs of coal and NG over the planning period are taken from the EIA Annual Energy Outlook 2016 (see Figure A.3) (EIA, 2016). A full description of the cost and technological assumptions for the existing and new generator clusters can be found in Table A.10 and Table A.11, respectively.

A summary of key system parameters used in both GEP models is presented in Table 4.3. The spinning and operating reserve requirements, implemented for each representative time block (hour or season) modeled in both GEP models, are based on parameters reported in the ReEDs model (Short et al., 2011). The planning reserve margin of 13.75% corresponds to the value used by ERCOT as part of its annual resource adequacy assessments. The annual load growth rate is derived from the forecast developed by ERCOT (ERCOT, 2015b). The discount rate is set to be equal to the nominal value of weighted average cost of capital assumed in the NREL annual technology baseline (ERCOT, 2016a). In this study, we consider one unique discount rate for all technologies, as is commonly implemented for GEP studies. However, in practice, this need not be the case. It is also worth noting that the assumed capital costs of individual generator types, sourced from the NREL annual technology baseline (Figure A.1), considers the unique cost of construction period financing for each technology.

We develop the load and renewable energy data for both GEP models based on the historical hourly profiles for ERCOT during the period 2004-2010 (ERCOT, 2016a, 2017) and the methodological approaches described in Section 4.2.1 and 4.2.2. Because the load

Table 4.3: Summary of key system parameters used in the TS-GEP and C-GEP

System parameters	Value	Reference
Spinning reserve requirement (% of load)	3%	(Short et al., 2011)
Total operating reserve requirement (% of load)	7.5%	(Short et al., 2011)
Planning reserve margin (% of peak load)	13.75%	(Potomac Economics, 2016)
Annual load growth rate	1.4%	(ERCOT, 2015b)
Discount rate	5.4%	(ERCOT, 2016a)

profiles of both GEP models are sampled from historical data, the extremes in their load duration curves are less pronounced than what is present in the actual data as shown in Figure 4.3A, although the load representation in the C-GEP is a better approximation of the historical data than the TS-GEP. Similar trends are observed when comparing the sampled data sets and historical data for renewable energy technology capacity factors (4.3B-C). In particular, Figure 4.3C illustrates how the time slice representation based on averaging time series over seasonal time blocks results in a poor characterization of the extreme observations in the historical wind capacity factor data (low and high). It is also worth noting that despite the differing temporal resolutions of both GEP models, the average annual capacity factors of all renewable energy technologies modeled are very similar (Table A.14). For example, the annual average capacity factor for new wind generators modeled in TS-GEP and C-GEP are 39.2% and 38.9%, respectively. In both GEP models, we model renewable energy generation based on a capacity factor time series that remains constant from one planning year to the next within the model time horizon, while we model the load as a time series with the same variability, but increasing annual average to reflect the assumed annual load growth of 1.4% per year (Table 4.3).

4.2.2 Selecting time slices for the TS-GEP

Since we are interested in comparing the outputs of two GEP models with differing temporal resolution of grid operations, the method of selecting the sampled load and renewable

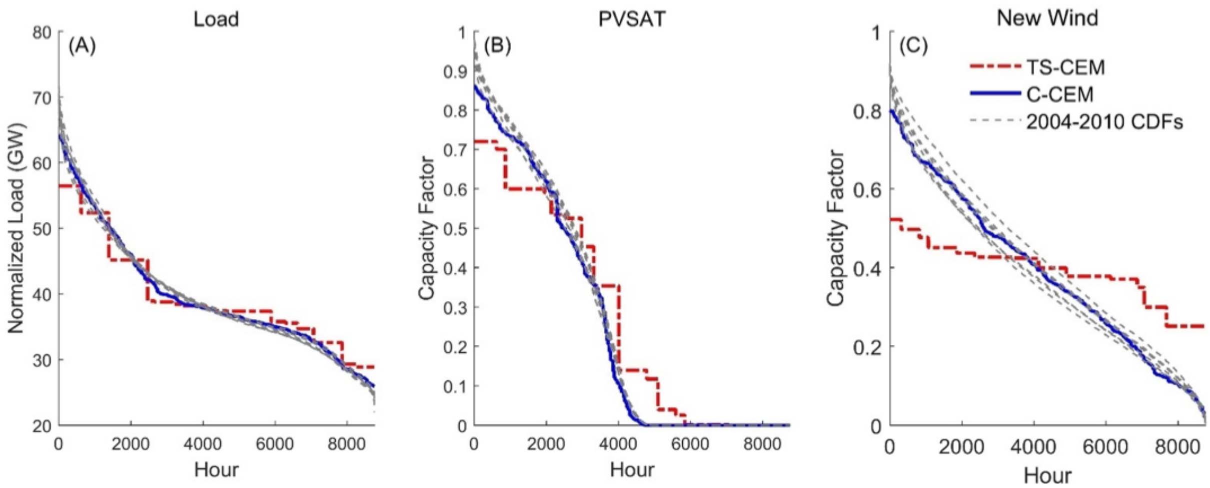

Figure 4.3: Load and renewable energy capacity factor duration curves comparison between historical data (2004-2010, shown in gray) and the corresponding duration curves assumed in the TS-GEP and C-GEP. A) Load duration curve comparison after the historical load data for all years is adjusted to have the same mean as observed in 2014. The mean absolute error (MAE) between the duration curves is smaller for the C-GEP (MAE = 694 MW) compared to the TS-GEP (MAE = 1941 MW). B) PV with single axis tracking (PVSAT) capacity factor duration curves result in an MAE of 0.015 and 0.056 for the C-GEP and TS-GEP, respectively. C) New Wind capacity factor duration curves result in an MAE of 0.022 and 0.125 for the C-GEP and TS-GEP, respectively, again indicating that the C-GEP better approximates the historical curves than the TS-GEP.

energy capacity factor data is essential to the analysis. For the TS-GEP, we construct time slices by first clustering consecutive days into "seasons" (loosely corresponding to spring, summer, fall, and winter). Within each resulting season, four time slices were created using the hourly clusters as proposed in ReEDS (Short et al., 2011): morning (7 am - 2 pm), afternoon (2 - 6 pm), evening (6 - 11 pm), night (11 pm - 7 am). Note that ReEDS considers a 17^{th} time slice to capture the 40 peak load days in a year.

We determined seasons for the TS-GEP using aggregate ERCOT load data from 2004-2010 while excluding leap days. As shown in Figure 4.4, aggregate load for each year was normalized between 0 and 1. A variance-minimizing clustering method was used to determine the four seasons and works as follows: The year was partitioned into 4 clusters by specifying 4 values marking the "start week" of each season. The same "start week" was used for each of the seven historical years of data. Given clusters, the variance within each cluster was computed using each day (a 24-hour vector) as a sample point/observation. To determine the optimal clusters, i.e. the best “start weeks,” we iterated over all possible “start weeks” for each season, such that no two seasons overlapped, and identified the cluster with the smallest variance. For example, the start week for summer was chosen from weeks 20-48. This approach is arguably more systematic than the procedure used in ReEDS where seasons are determined simply by grouping consecutive months (Short et al., 2011). Once time slices were determined, renewable energy capacity factors in each time slice were computed as the average capacity factor over all hours in each time slice.

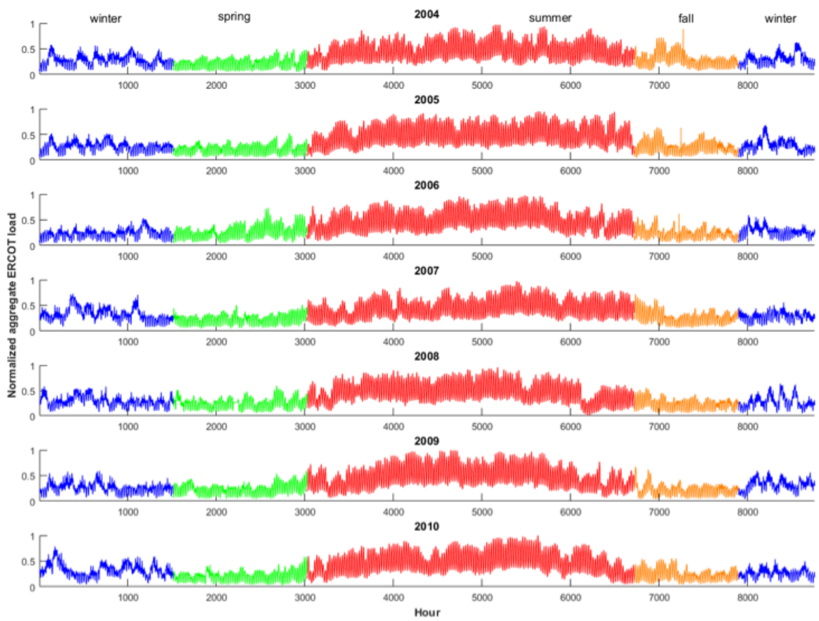

Figure 4.4: Seasons used in the TS-GEP, defined using a variance minimizing clustering method applied to seven years of historical load data (normalized between 0 and 1). The “start week” of each season is: spring = week 9, summer = week 18, fall = week 40, winter = week 47.

4.2.3 Selecting representative days for the Chronological GEP

For the C-GEP, we utilize a k-means clustering procedure to determine representative days for modeling annual grid operations. The goal of the clustering procedure is to select representative days that closely approximate (i) the cumulative distribution functions (also known simply as “duration curves”) of historical load and renewable energy time series, (ii) the temporal correlation of each time series, and (iii) the hourly correlation between load and renewable energy time series. The first aim ensures that annual load and renewable energy capacity factors are adequately represented. The second attempts to ensure that inter-hour variability is characteristic of the actual system, and thus adequately captures the need for increased ramping in certain hours of the day. The third attempts to ensure that load and renewable energy profiles sufficiently characterize the correlation between these time series throughout the year. The data set used for clustering was the hourly ERCOT data (ERCOT, 2016a, 2017) for the seven year period from 2004 to 2010, with leap days excluded. Let $D = \{1, ..., D = 2555\}$ denote the set of days for the seven year period from 2004 to 2010, let $H = \{1, ..., 24\}$ denote the set of hours in a day, and let $K = \{\text{load}, \text{csp}, \text{pvsat}, \text{old wind}, \text{new wind}\}$ denote the set of load/technology types considered for clustering. Note that we assume the capacity factor profile for new and existing PV plants to be identical, partly because of data availability and the small amount of installed PV capacity in ERCOT as of 2015 (Table 4.2). In what follows, a "point" denotes a vector or time series of data associated with a given day. Specifically, let $\boldsymbol{x}_d = (x_{dhk})_{h \in H, k \in K'}$ denote a vector of hourly data associated with a subset of types in K'. For example, if $K' = \text{load}$, then x_{44} denotes a vector of hourly load data, for all of ERCOT, corresponding to the 44^{th} day in the data set. Several approaches for selecting representative days for long-term power systems expansion models have been considered. de Sisternes and Webster (2013) introduce an approach for optimally selecting sample weeks to approximate net load for long-term generation planning problems. Poncelet et al. (2015) present a mixed-integer linear optimization approach to simultaneously address the three objectives listed at the outset. Nahmmacher et al. (2016a) propose an agglomerative clustering algorithm that begins with D clusters, each consisting of exactly one of the original D observations (days) in the data set. The two "closest" clusters are then merged reducing the number of clusters by 1. This procedure continues until a single cluster consisting of all D observations remains.

Since we ultimately selected a k-means approach over a hierarchical approach, it is worth making some qualitative remarks about the latter. Agglomerative clustering is known to perform well (often better than centroid-based methods like k-means) when the underlying data consists of multiple disjoint "islands." This superior performance occurs because there will eventually be an iteration in which two dissimilar "islanded" clusters are deemed "closer" to one another than all other cluster pairs, and merging these clusters will result in a significant increase in the intra-cluster variance for the new cluster. In the context of power systems planning where observations are historical load and capacity factor profiles, one is tempted to claim that the data naturally decomposes into easily separable "islands." In our experience, this was not the case. For example, while there are clearly days (observations) with a single peak load and other days with multiple peaks, there were many days that possess characteristics of both profiles (see Figure 4.5). As a consequence, there are few iterations in which the intra-cluster variance significantly "jumps." Worse, an agglomerative clustering algorithm often produced a single large cluster with many observations and large intra-cluster variance, along with many small clusters with only a handful of observations.

In contrast, centroid-based clustering is much more blunt in forming clusters, meaning it assumes that the underlying data come from spherical (Gaussian distributed) clusters. Below, we provide a step-by-step description of our approach along the lines of what was done in Nahmmacher et al. (2016a).

Step 1: Normalizing all time series and selecting a distance metric

Clustering algorithms attempt to group similar observations into the same cluster. Fundamental to any clustering method is the choice of distance metric used to quantify the degree of similarity between two observations. Indeed, the importance of choosing an appropriate distance metric is often under-emphasized and/or poorly understood. Because most off-the-shelf algorithms have a set of pre-defined distance metrics (e.g. L1, L2, and cosine), most practitioners decompose the distance metric selection problem into two problems: data normalization and selecting a pre-defined distance metric. We have chosen this approach as well. Together, the choice of normalization and distance metric have a significant impact on the ultimate clusters. We normalize all load data between 0 and 2 for each year. Renewable energy capacity factors were already normalized between 0 and 1, due to their inherent

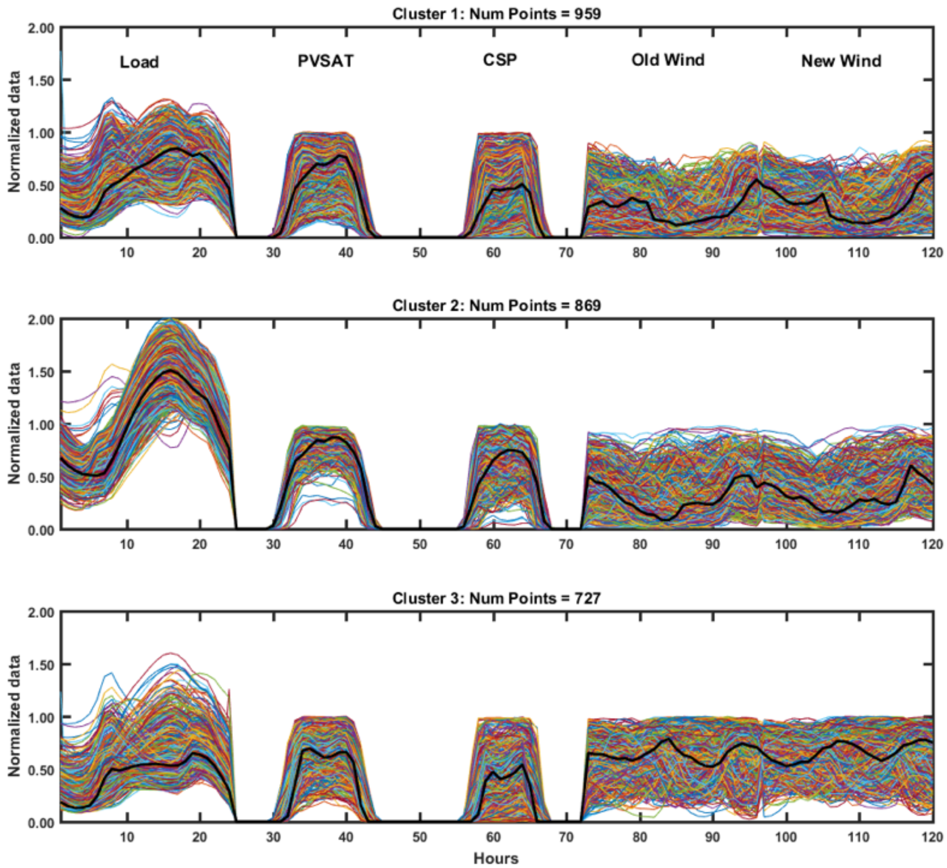

Figure 4.5: Clusters produced by k-means algorithm when k=3 clusters using aggregate ERCOT data, 100 replications, and the L2 distance metric. Five times series – Load, PVSAT, CSP, Old Wind, and New Wind – are "stitched together" so that each historical day is stored as a single 120-dimensional vector. Load is normalized between 0 and 2, whereas all renewable energy capacity factors are normalized between 0 and 1. Each subplot depicts the points/days in the cluster (shown in color) and then most representative day (shown as a single black line). The title of each subplot indicates the number of points/days assigned to that cluster.

nature. The normalization is shown in each subplot in Figure 4.5 where load, the first 24 components of the 120-sized vector, is normalized between 0 and 2, and the renewable energy capacity factors are normalized between 0 and 1 in the remaining 96 ($= 24 \times 4$) components. We test this normalization scheme with two distance metrics – L1 and L2 norms – and use the L2 norm method to develop clusters for the C-GEP. As part of a sensitivity analysis, we present the outputs of the C-GEP using the L1 norm based clustering approach in Section A.3.

Step 2: Applying the clustering algorithm and deriving a candidate set of clusters

Clusters were determined using the “kmeans” function in MATLAB with 100 replications. Because our interest is to select a small number of representative days to include in the model, we called “kmeans” R times, for values $r = 1, ..., R = 12$, producing results with 1 to R clusters.

Step 3: Choosing one representative day per cluster

For each cluster, we select the historical day closest (using the a priori selected distance metric) to the centroid of that cluster as the most representative day. This is different from using the centroid of the cluster, which may not capture the true variability seen within a day. Thus, the most representative days are indeed historical days. Figure 4.5 shows the most representative day selected for each cluster when three clusters are used.

Step 4: Weighting each representative day according to its cluster size

Each representative day is assigned a weight proportional to the number of historical days in the corresponding cluster. Specifically, given a total of $D = 2555$ days in the data set and D_c days in cluster c, the weight assigned to each representative day is $w_c = D_c/D$. For example, cluster 1 in Figure 4.5 possesses 959 days and receives a weight of 959/2555.

Step 5: Scaling single time series in order to reach the correct annual average

Finally, the weighted load profile was normalized to equal 2015 aggregate ERCOT load of 347.5 TWh so that a fair comparison between runs with different numbers of clusters could be made. Let x_c^* denote the historical days selected as the most representative days. Then, the data is scaled such that:

$$\sum_c w_c \sum_h x_(c, h, Load)^* = 347.5TWh \tag{4.1}$$

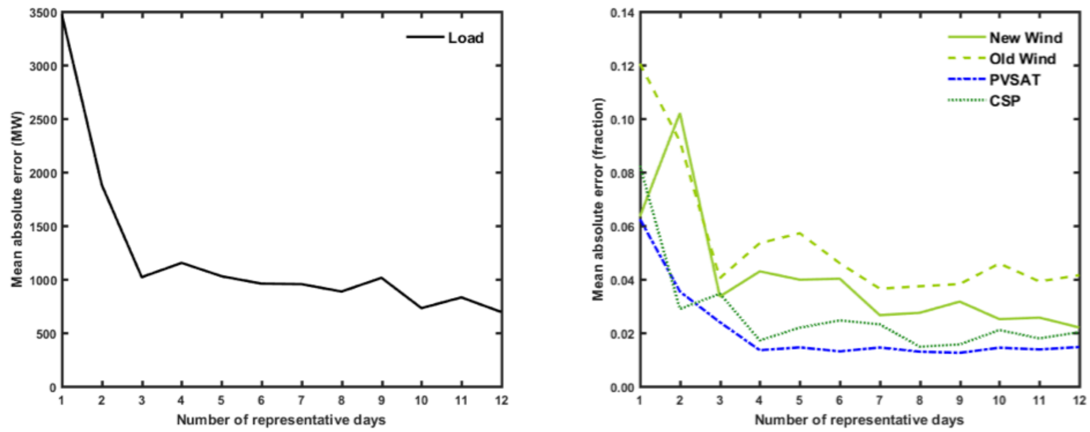

Figure 4.6: Mean absolute error in the load duration curve (left panel) and the cumulative distribution curves of renewable capacity factors (right panel), relative to the corresponding historical ERCOT curves, for a varying number of representative clusters

Figure 4.6 presents the approximation accuracy improvement in the load duration curve and cumulative distribution curves due to the above clustering procedure, for an increasing number of representative days. Specifically, it shows the mean absolute error (expressed in MW) in the load duration curve and the mean absolute error (expressed as a fraction) in the cumulative distribution curves of the renewable energy resources relative to seven years (2004-2010) of hourly ERCOT data. As expected, the error tends to decrease as more representative days (more clusters) are included. There are at least two reasons why the curves are not monotonically decreasing. First, the day chosen as "most representative" is not the centroid of the cluster, but the one closest in Euclidean distance to the centroid. Second, to determine the clusters, hourly profiles for load and all renewable energy technologies are considered simultaneously, not individually. Thus, while the error in the aggregate profile (the 120-sized vector including all technologies simultaneously) declines nearly monotonically as the number of clusters increases, the individual technologies do not exhibit this trend. On the other hand, Figure 4.6 shows that the error in two technologies tend to offset each other. For example, given two clusters, the sudden decrease in error for CSP is met with a simultaneous rise in error for New Wind. With three clusters, the opposite occurs: error in New Wind decreases, whereas CSP error increases. Note that the mean absolute error in the

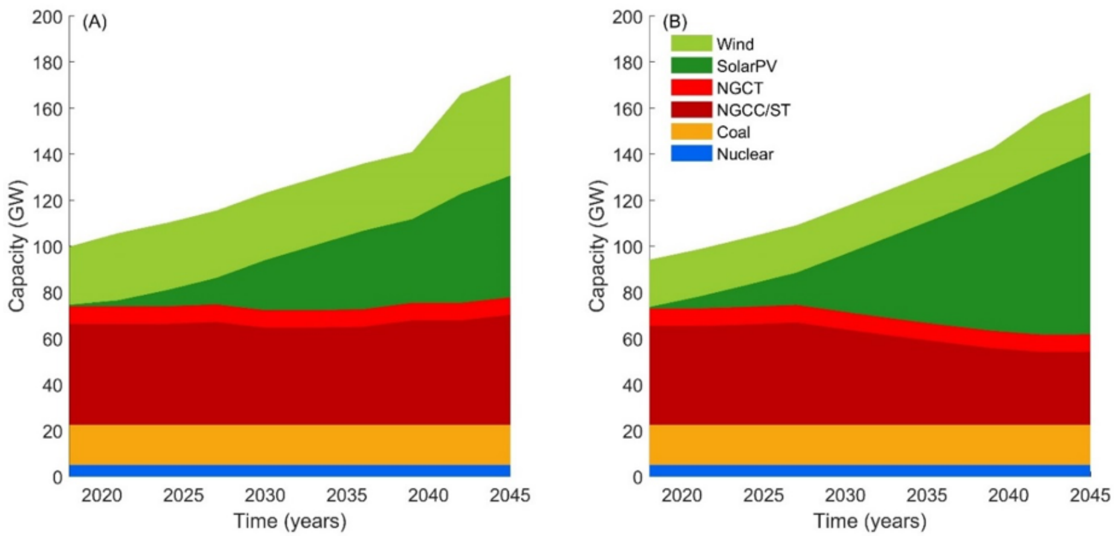

Figure 4.7: Comparison of capacity projections by (A) the chronological GEP (C-GEP) and (B) the time slice GEP (TS-GEP) in the 50% renewables case.

cumulative distribution curves used for the TS-GEP are 1941 MW for load, 0.125 for New Wind, 0.132 for Old Wind, 0.056 for PVSAT, and 0.071 for CSP. These values are close to the errors found for a 1-representative day cluster, but otherwise uniformly larger than any other choice of number of clusters.

4.3 Results

4.3.1 Comparing GEP models with differing temporal resolution

Figure 4.7 presents projections of capacity (2015-2045) for a hypothetical 50% renewables scenario, estimated by the C-GEP and TS-GEP using a consistent set of input assumptions described in the prior section and Appendix C.8. In this scenario, a target of 50% renewable energy penetration is set for model years 2040 and beyond. The main drivers for capacity additions in both models are the assumed growth in electricity demand of 1.4% per year over the planning horizon (ERCOT, 2015b) and the imposed renewable energy penetration target in 2040 and beyond. Both GEP models project approximately 175 GW of grid capacity in 2045, where solar PV dominates new installations. However, the magnitude of solar PV installed varies in both models, with the C-GEP demonstrating a preference for a portfolio of options including wind and natural gas combined cycle (NGCC) capacity. Figure A.4 demonstrates a similar preference for solar PV in the TS-GEP annual generation mix over the planning period. Both models also project that some existing natural gas (NG) capacity will be retired – mainly older natural gas steam turbines (NGST) in both models and some additional natural gas combined cycle (NGCC) plants in the TS-GEP. The TS-GEP's time slice representation of load and capacity factors for renewable energy generation overlooks the hour-to-hour variability in load and capacity factors within these time blocks and the limited ability of thermal generators to adjust their output accordingly. In contrast, the C-GEP better approximates the observed hourly variability in load and renewables and includes explicit limits on the flexibility of thermal generator fleet, through hourly ramping limits and startup costs. The C-GEP therefore projects the need for adding flexible thermal capacity, mostly as new NGCC plants, and projects lower solar PV penetration than the outputs of the TS-GEP. In addition, the variability in wind capacity factors may not be well characterized by a time slice representation (see comparison of time slice sampling results and historical data in Figure 4.3) (Bistline et al., 2017; Blanford et al., 2018). This may partly explain why the TS-GEP results undervalue wind capacity additions compared to the C-GEP, even though the average annual capacity factors of new wind plants are similar in both GEP models (Table A.14).

In addition to the 50% renewables scenario, we analyze a range of hypothetical scenarios to evaluate how the outputs of both GEP models change with increasing renewable energy penetration. We test the two GEP models by constraining them to meet a specific percentage (40% - 70%) of dispatched generation by 2040 (and beyond), resulting in a total of 4 renewable energy scenarios[2]. We compare the capacity mix at the end of the planning period across the 4 renewable energy scenarios between both models. The results, presented in Figure 4.8A (reference scenario set), suggest that the TS-GEP shows a strong preference for solar PV installations over wind and NG compared to the C-GEP for a majority of

[2]These renewable energy targets were chosen to exercise and ultimately contrast the two expansion models and do not reflect the authors' opinion or endorsement that such targets are economically viable or attainable.

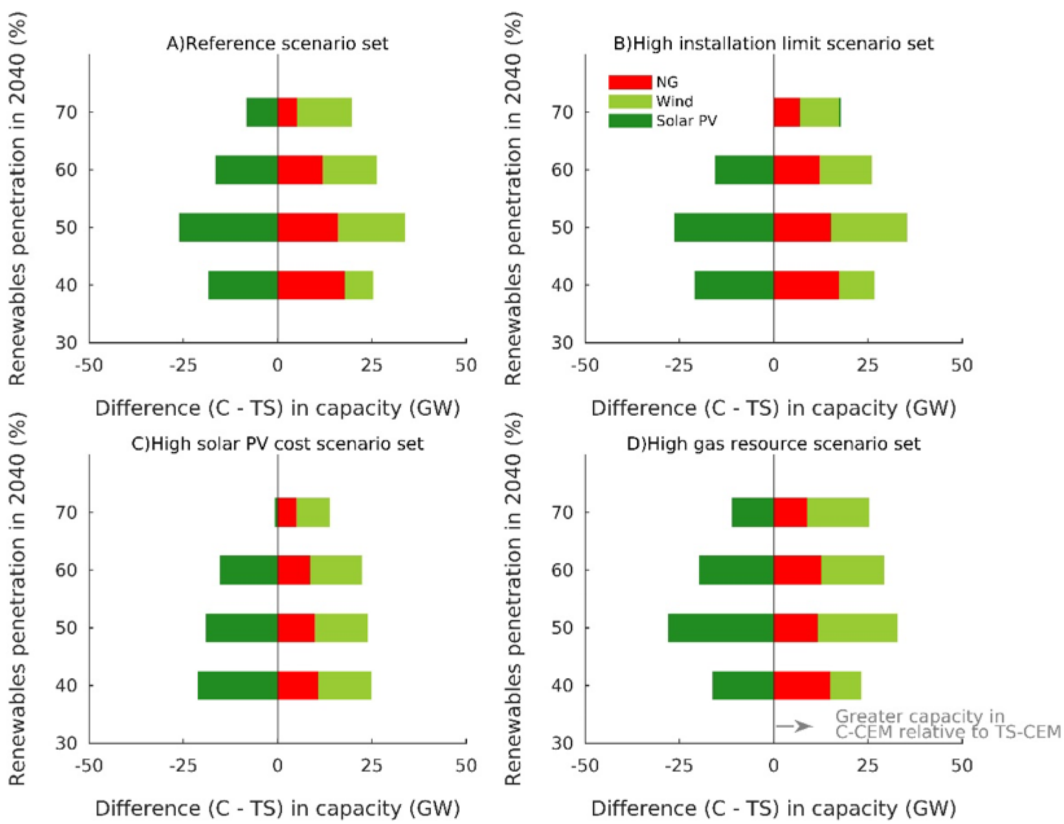

Figure 4.8: Difference in 2045 capacity projections for solar PV, wind and natural gas (NG) in the chronological GEP (C-GEP) compared to the time slice GEP (TS-GEP) for a range of renewable energy scenarios up to 70%, and across 4 alternate scenario sets. Here "NG" includes all types of natural gas plants including NGCC, NGCT and NGST.

scenarios. For instance, in the 40% renewables scenario, the C-GEP projects only 50 GW of solar PV by 2045, while the TS-GEP projects 70 GW (35% higher). As the renewable energy penetration target is increased, however, the difference in PV capacity estimated by the TS-GEP and the C-GEP tends to decrease. This trend is partly an outcome of the two models diverging in their estimates of curtailed renewable energy generation with increasing renewable penetration[3] (see Figure 4.9 and Figure A.5). For example, for the 70% renewables scenario, the C-GEP projects 11% of renewable energy generation is curtailed in 2045 as

[3]The C-GEP's increased granularity of representing hourly grid operations, including ramping limits and on/off status of individual thermal generators, partly explains why it estimates increasing curtailment with increasing renewable penetration. In contrast, the curtailment observed for the TS-GEP tends to plateau with increasing renewable penetration, likely because it does not capture the hour-to-hour variability in load and renewable output and overlooks on/off commitment of all thermal generators (see Table 4.1).

compared to 6% estimated by the TS-GEP. All else equal, higher curtailment implies that greater wind and solar PV capacity is required to dispatch the same amount of renewable energy generation. Under higher renewable penetration scenarios (e.g. 70% renewables), this effect counterbalances the preferential additions of PV over wind in the TS-GEP due to its lower temporal resolution, resulting in a diminished PV capacity difference between the model outputs. In higher renewable energy scenarios (60% and 70%), some coal (2-5 GW) and nuclear (<3 GW) retirements are also seen in the TS-GEP, which are not observed in the C-GEP (not shown in Figure 4.8). The TS-GEP's minimum turndown constraints on coal and nuclear plants (Table 4.1) result in the inability of these plants to cycle to the same extent as in the C-GEP, which explains why the TS-GEP chooses to retire some coal and nuclear capacity under higher renewable energy scenarios.

To evaluate how robust the aforementioned differences in the projected capacity mix of the TS-GEP and C-GEP are to various technology and cost assumptions, we compare GEP outputs for three additional scenario sets: B) doubling the annual installation limits on wind and solar PV compared to the reference scenario set (see reference data in Table A.12); C) higher solar PV cost compared to projections used in the reference scenario set (presented in Figure A.2); and D) a higher gas resource scenario (i.e. lower gas price), also from the EIA Annual Energy Outlook 2016 (EIA, 2016) (presented in Figure A.3). Results across all scenario sets confirm the same trends observed in the reference scenario set: the TS-GEP shows a strong preference for solar PV installations over wind and NG, compared to the C-GEP across a majority of the evaluated scenarios, with reduced differences seen at higher renewable energy scenarios. Notably, in the high solar PV cost scenario, both models are directly dis-incentivized to build solar PV and therefore the differences between the model outputs are relatively smaller (Figure 4.8C).

4.3.2 Using production cost simulations to assess GEP results

Annual hourly grid operations are approximated in both GEP models, although in different ways. To test how robust these approximations are, we solve the PCS model to simulate grid operations for a single year (2045) with hourly resolution, using the capacity projected by both GEP models. To account for inter-annual variability in load and renewable energy generation, we solve the PCS model for seven different realizations of time series for load and renewables capacity factors, for each renewable energy scenario. The data for the load and renewables capacity factor profiles were derived from historical data (2004-2010) available from ERCOT (ERCOT, 2016a, 2017). We then compare the outputs of the GEP models and the PCS model for each renewable energy scenario using a range of metrics, including annual curtailment (Figure 4.9A), unmet demand (Figure 4.9B), annual renewables penetration (Figure A.6), and annual thermal generation mix (Figure A.7).

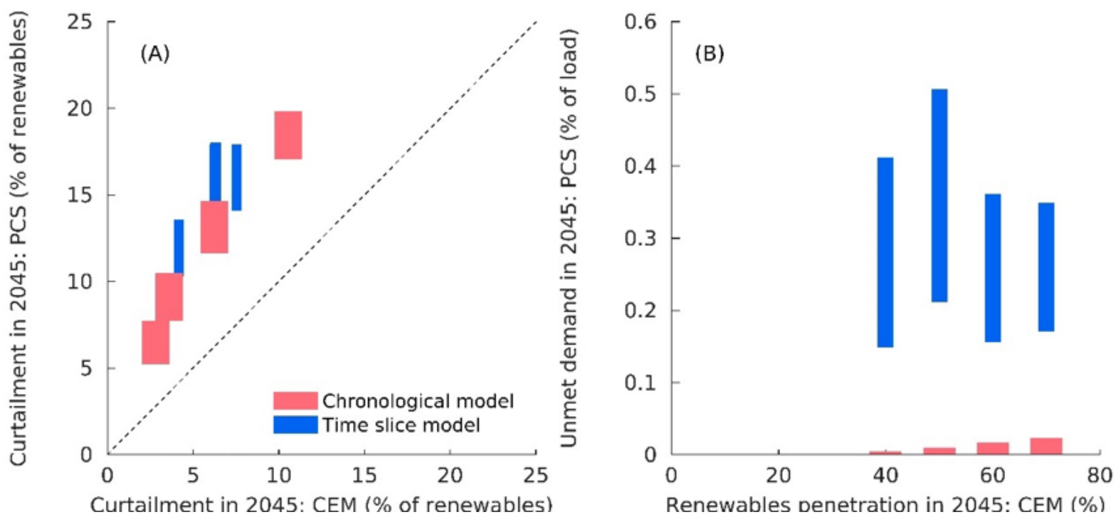

Figure 4.9: (A) Comparison of 2045 curtailment in the GEP models to curtailment in the Production Cost Simulation (PCS) model for the same capacity mix. (B) Comparison of unmet demand in 2045 in the PCS model for the same capacity mix as projected by the GEP models. The GEP outputs correspond to the reference scenario set (Figure 4.2A). The height of each bar corresponds to the range of values obtained from the PCS model by simulating seven different realizations of time series for load and capacity factors for renewable energy generation (further details on load and renewable energy data available in Section 4.2).

In Figure 4.9A, the individual bars refer to the 4 renewable energy scenarios presented earlier and the height of each bar represents the range of outputs from the PCS model when considering seven different realizations of load and renewable energy generation time series. The inability of either GEP to accurately capture extreme situations such as maximum or minimum net load (i.e. load minus renewable energy generation), maximum renewable energy generation or rapid changes in renewable energy output partly explains why both GEP models underestimate curtailment. For the C-GEP capacity mix, curtailment projected by the PCS generally tends to increase as the prescribed renewables penetration level increases. This is not surprising since both the C-GEP and the PCS have similar thermal generator operating constraints that directly contribute to instances of curtailment.

Both GEP models estimate that demand is met for all the representative time steps included in the models. In contrast, when the capacity mix estimated by these models is input to the PCS model, demand remains unmet in some hours. This is expected, since both GEP models approximate grid operations and do not capture the full extent of variability of grid conditions. As seen in Figure 4.9B, increasing the model resolution (C-GEP) as compared to seasonal-average modeling (TS-GEP) certainly reduces instances of unmet demand. It is also worth noting that, although the unmet demand for the C-GEP capacity mix increases with increasing renewable energy targets, the maximum value is still relatively small at 0.023% of load for the 70% renewables scenario, which is comparable to the loss of load threshold values considered in estimating reserve margin requirements (Pfeifenberger et al., 2013). Finally, the C-GEP outputs also better approximate the annual generation mix by technology type, when compared to the TS-GEP outputs, as shown in Figure A.6 and Figure A.7.

4.3.3 Impact of the number of representative days within the C-GEP

Within the C-GEP, the number of sample days selected to represent the entire year's load and renewable energy generation may also impact results, including capacity and generation projections. We assessed this aspect by solving the C-GEP using a range of sample days (1 day up to 12 days). Such a question has been considered in the context of a linear GEP by Nahmmacher et al. (2016a), who concluded that a GEP with fewer representative days selected to represent annual grid operations may result in higher projections of renewable energy capacity. In our experiments using the C-GEP, which is a MILP model, we use a k-means clustering approach with the L2-norm used as the distance metric, to select these days from the historical data, with the total annual load adjusted to be identical across all 12 scenarios. Figure 4.4 compares the resulting capacity projections for NG, solar PV and wind, for each of the scenarios with a different number of sample days, under a 50% renewable energy penetration target (comparison of generation shown in Figure A.8). Although the total renewable energy capacity remains nearly the same across all scenarios, the projected PV capacity and wind capacity have increasing and decreasing trends, respectively, with increasing number of representative days selected. For instance, the solar capacity in the 1-day scenario is approximately 20% higher than the capacity for the scenario with 12 representative days. The NG capacity also follows an increasing trend with increasing representative days, with up to 10% differences between the two bookend cases in Figure 4.10. The non-monotonic trends in the installed capacity of individual technologies in Figure 4.4 are partly an outcome of the approach for selecting representative days. Specifically, as discussed in Section 4.2.3, historical data are clustered according to the joint distribution of load and renewable energy capacity factors, which does not guarantee that the error in representing individual distributions (capacity factor or load) will exhibit a monotonically decreasing trend. The capacity trends observed in Figure 4.10 were also found to be robust to changing the distance metric used in the selection of representative days from the L2-norm to the L1-norm (see Figure A.9 and Figure A.10). Overall, the results are consistent with the trends reported when comparing the GEP models: with lower temporal resolution (number of sample days selected, in this case), solar PV capacity is overestimated while wind and NG capacity are underestimated relative to the higher resolution GEP models.

For long-term energy planning models like NEMS or ReEDS with multi-sector scope or large spatial coverage or both, computational issues may make it impractical to include 12 representative days in the power system expansion model. In such cases, representing annual grid operations using fewer representative days could still provide some improvement over a traditional time slice approach, with regard to addressing the variability of load and renewable energy generation and other grid operating constraints. This point is illustrated in Figure 4.10, where for instance, the solar, wind and NG capacity projected using 6 repre-

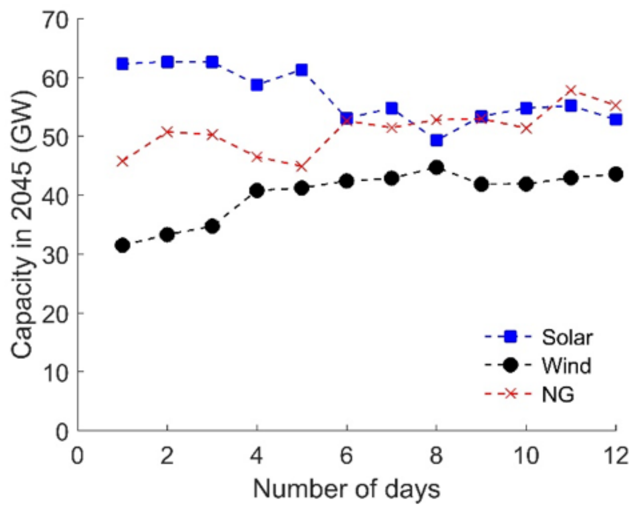

Figure 4.10: Capacity projections for solar, wind and natural gas (NG) in 2045, using the chronological GEP (C-GEP) under a 50% renewable energy scenario, varying as a function of the number of sample days selected to represent load and renewables data for annual grid operations. (The L2-norm is used in the k-means clustering approach for choosing the representative days)

sentative days is within 5% of the total installed capacities projected by the GEP using 12 representative days.

4.4 Conclusions

In this chapter, we perform a systematic comparison of two alternate GEP frameworks to demonstrate how the choice of representing grid operations within a power system GEP framework can impact future projections of grid evolution. For the same set of technology and cost assumptions, we find that a GEP with time slice representation of grid operations (e.g. the TS-GEP developed here) is in general likely to overestimate solar PV capacity (by 35% in one case) and underestimate wind and the supporting NG capacity requirements, compared to a GEP with higher temporal resolution and generator ramping and startup constraints (C-GEP). This finding is explained primarily by the limited representation of the temporal variability in renewable energy generation, notably wind, and its correlation with load when using the time slice approach as compared to the chronological approach. For solar PV, using values of capacity factors based on 4-hr seasonal averages (as in the

TS-GEP) overvalues the coincidence between peak solar PV generation and peak system load (also a 4-hr seasonal average) and consequently underestimates the declining value of solar PV generation with increasing penetration, as compared to the chronological approach using 12 representative days (as in the C-GEP) at an hourly resolution. The differences in the capacity mix to achieve the same renewable energy targets have reliability implications, as reflected by the lower unmet demand projected for the C-GEP capacity mix when tested in a detailed hourly simulation of annual grid operations.

While it is common for policy-focused GEP studies to test the capacity mix estimated by a GEP through a PCS framework (Brinkman, 2015; Lew et al., 2013), our study highlights the importance of evaluating the operational performance of the capacity mix projections for multiple years of load and renewable energy generation profiles. Such an analysis benchmarks the ability of the capacity mix to achieve the desired reliability and/or environmental attributes. For instance, the results presented here suggest that the unmet demand resulting from the capacity mix estimated by the C-GEP (using 12 representative days) is less sensitive to the annual variations in load and renewable energy generation profiles compared to the outputs projected by the TS-GEP (Figure 4.9B).

Even within a C-GEP framework, selecting fewer than 4 sample days may lead to considerable overestimation of solar PV capacity. This finding has implications for the choice of temporal resolution in not just power sector planning models, but also more broadly for multi-sector, multi-country energy economic and integrated assessment models. For example, it was recently suggested that the current time slice implementation in the electricity grid planning implementation of the 2016 NEMS energy-economic model for the US may be overestimating solar PV capacity projections (Wood, 2016). Similarly, Bistline et al. (2017) performed an intra-model comparison of alternative temporal representations in the US-REGEN model and concluded that using a seasonal-average approach (akin to TS-GEP) is likely to overstate renewables capacity and understate investment in dispatchable generation, compared to the representative hours approach (akin to C-GEP). Our study contributes to the growing body of evidence on the need for using a temporal representation based on a few representative days or other parameterizations that yield similar behavior in multi-sector, energy-economic models and other energy system models supporting policy analysis and decision-making. There are several future research directions worth investigating. While this study focused on the effect of changing temporal resolution in a GEP while keeping the spatial resolution constant, it would be interesting to consider the relative importance of spatial and temporal resolution by repeating the analysis in the context of a GEP co-optimizing generation and transmission expansion. Given the growth in energy storage technologies, it would be instructive to understand the impact of energy storage on renewable energy penetration projections in future electricity grids, as well as the necessary temporal resolution required to adequately account for their attributes in a GEP model. Note that we have developed a GEP similar to the C-GEP that considers energy storage and is presented in chapter 5. Since parameter uncertainty (e.g., in construction lead times and capex costs) in GEP models is always a prominent issue, it could be valuable to implement stochastic versions of our models and perform a similar analysis. Finally, it would be interesting to consider within a GEP framework, the trade-offs between deploying storage, solar PV systems at the centralized and distributed scale, given the different types of grid services available from deployment at each scale.

Chapter 5

Deterministic Electric Power Infrastructure Planning: Mixed-integer Programming Model and Nested Decomposition Algorithm

In this chapter, we propose an optimization modeling framework to evaluate the changes in the power systems infrastructure required to meet the projected electricity demand over the next few decades, while taking into account detailed operating constraints, and the variability and intermittency of renewable generation sources. The modeling framework, which is based on mixed-integer linear programming (MILP), takes the viewpoint of a central planning entity whose goal is to identify the source (nuclear, coal, natural gas, wind and solar), generation technology (e.g., steam, combustion and wind turbines, photovoltaic and concentrated solar panels), location (regions), and capacity of future power generation technologies that can meet the projected electricity demand, while minimizing the amortized capital investment of all new generating units, the operating costs of both new and existing units, and corresponding environmental costs (e.g. carbon tax and renewable generation quota).

The major challenge lies in the multi-scale integration of detailed operation decisions at the hourly (or sub-hourly) level with investment planning decisions over a few decades. In order to improve its computational tractability, judicious modeling approximations and aggregations are considered. The major contribution of this chapter is the combination of formulation, solution strategy, and application to a case study based on real-world data. We develop a decomposition algorithm based on Nested Benders Decomposition for mixed-integer multi-period problems to solve large-scale models. This framework was originally developed for stochastic programming by Zou et al. (2018b), but we have adapted it to deterministic multi-period problems. We have modified it to handle integer and continuous state variables, at the expense of losing the finite convergence property due to potential duality gap, and have applied acceleration techniques to improve the overall performance of the algorithm.

In Section 5.1, a formal problem statement is given, and in Section 5.2 we describe the modeling strategies adopted to handle the spatial and temporal multi-scale aspect of the problem. The MILP formulation is presented in Section 5.3. Section 5.4 describes the proposed decomposition algorithm. Section 5.5 shows the results for a real-world case study for the Electric Reliability Council of Texas (ERCOT) region, and a comparison between the performance of the full size MILP formulation and the proposed algorithm.

5.1 Problem statement

The proposed planning problem involves choosing the optimal investment strategy and operating schedule for the power system in order to meet the projected load demand over the time-horizon for each location.

A set of existing and potential generators is given, and for which the energy source (nuclear, coal, natural gas, wind or solar)[1] and the generation technology are known.

- For the existing generators we consider: (a) **coal:** steam turbine (coal-st-old); (b) **natural gas:** boiler plants with steam turbine (ng-st-old), combustion turbine (ng-

[1]In this thesis we do not consider hydroelectric power as it is available in very limited amounts in the ERCOT region.

ct-old), and combined-cycle (ng-cc-old); (c) **nuclear:** steam turbine (nuc-st-old); (d) **solar:** photovoltaic (pv-old); (e) **wind:** wind turbine (wind-old);

- For the potential generators we consider: (a) **coal:** without (coal-new) and with carbon capture (coal-ccs-new); (b) **natural gas:** combustion turbine (ng-ct-new), combined-cycle without (ng-cc-new) and with carbon capture (ng-cc-ccs-new); (c) **nuclear:** steam turbine (nuc-st-new); (d) **solar:** photovoltaic (pv-new) and concentrated solar panel (csp-new); (e) **wind:** wind turbine (wind-new);

Also known are: their nameplate (maximum) capacity; expected lifetime; fixed and variable operating costs; start-up cost (fixed and variable); cost for extending their lifetimes; CO_2 emission factor and carbon tax, if applicable; fuel price, if applicable; and operating characteristics such as ramp-up/ramp-down rates, operating limits, contribution to spinning and quick start fraction for thermal generators, and capacity factor for renewable generators.

For the case of existing generators, their age at the beginning of the time-horizon and location are also known. For the case of potential generators, the capital cost and the maximum yearly installation of each generation technology are also given. Also given is a set of potential storage units, with specified technology (e.g., lithium-ion, lead-acid, and flow batteries), capital cost, power rating, rated energy capacity, charge and discharge efficiency, and storage lifetime. Additionally, the projected load demand is given for each location, as well as the distance between locations, the transmission loss per mile, and the transmission line capacity between locations.

The problem is then to determine: a) location, year, type, and number of generators and storage units to install; b) when to retire generators and storage units; c) whether or not to extend the life of the generators that reached their expected lifetime; d) an approximate operating schedule for each installed generator; and e) the approximate power flow between each location in order to meet the projected load demand while minimizing the overall operating, investment, and environmental costs.

5.2 Modeling strategies and assumptions

The mix of combinatorial and operational elements of the problem described in Section 5.1 means that, depending on the time horizon and area considered, the corresponding optimization problem may be too large and intractable for current commercial general purpose MILP solvers. Therefore, in order to solve the resulting deterministic MILP to provable optimality for large areas and over a few decades, it is key to explore judicious modeling aggregations and approximations to address the multi-scale aspects, both in its spatial and temporal dimensions. In order to significantly reduce computation time, generator clustering (Palmintier and Webster, 2014) and time sampling (Pina et al., 2011) approaches are adopted.

5.2.1 Spatial representation

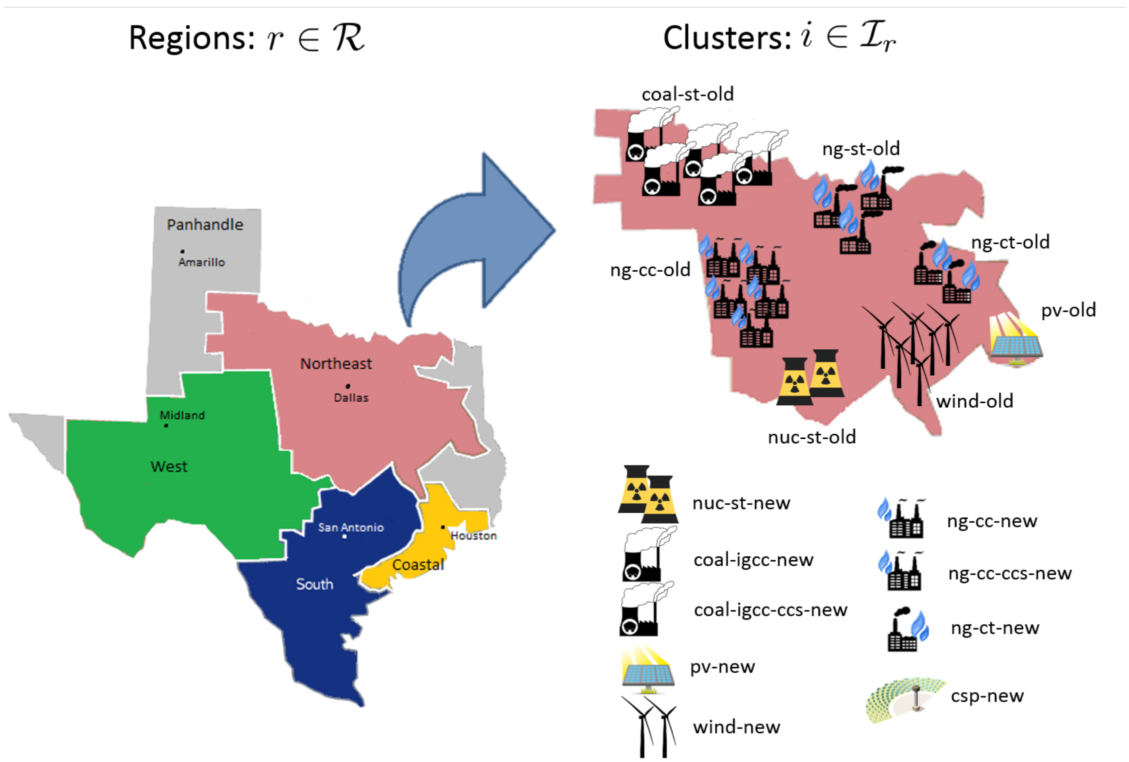

Figure 5.1: Model representation of regions and clusters (regional map modified from ERCOT (2016))

In order to allow the solution of large-scale instances, the area of scope is divided into regions that have similar climate (e.g., wind speed and solar incidence over time), and load demand profiles. It is assumed that the potential locations for the generators and storage units are the midpoints of each region r. Additionally, based on the work of Palmintier and Webster (2014), generators and storage units that have the same characteristics, such as technology and operating status (i.e., existing or potential), are aggregated into clusters i for each region r. The spatial configuration of the problem is shown in Figure 5.1 for the ERCOT region.

The major impact of this approximation in the model formulation is that the discrete variables associated with generators and storage units correspond to integer rather than binary variables to represent the number of generators/storage units under a specific status in cluster i.

5.2.2 Temporal representation

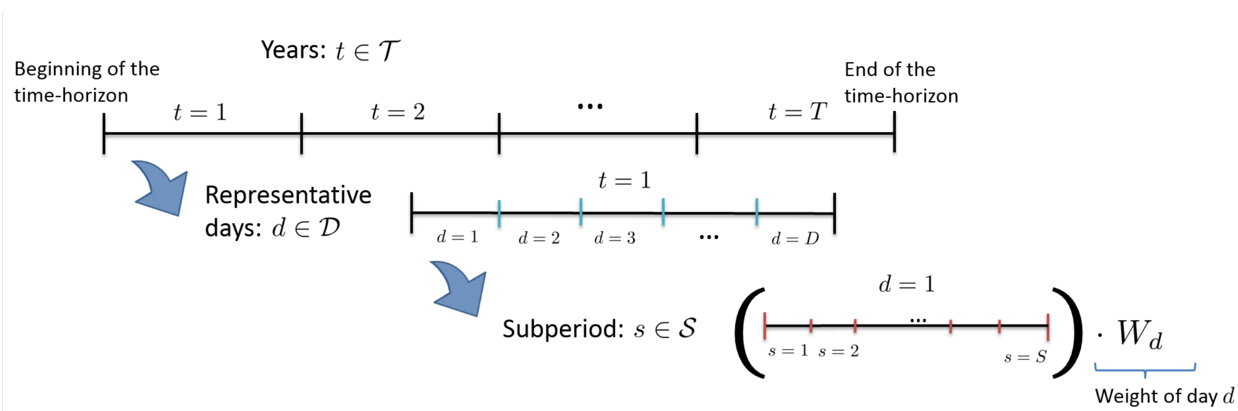

Figure 5.2: Multi-scale representation

It is crucial to include hourly level information to evaluate scenarios with increasing renewable energy generation (Pina et al., 2013), because of the variability in that resource, as well as the changes in load. On the other hand, strategic capacity expansion decisions must be optimized over a long-term horizon (e.g., a few decades). Therefore, the investment decisions are made on a yearly basis, while operating decisions are made at the hourly level. To tackle the problem's multi-scale nature and to reduce computation time, each year is modeled using d representative days with hourly resolution resulting in 24 subperiods. We employ a k-means clustering approach to select these days from historical data (see Section 5.5), where the goal of the clustering procedure is to select representative days to approximate: (i) the "duration curves" of historical load and renewables time series, (ii) the temporal correlation of each time series, and (iii) the hourly correlation between each time series. The temporal configuration of the problem is shown in Figure 5.2

5.2.3 Transmission representation

Transmission is also an important aspect of a power systems infrastructure, influencing where to build power plants, which ones to operate, and how much power to be generated by each of them. The rigorous way of representing transmission between generation and load nodes in the system is through optimal power flow models (e.g., Frank et al. (2012a,b)). As explained in Section 5.2.1, the proposed model uses a reduced network, which for the example in Figure 5.1 has only 5 nodes representing the 5 regions. Additionally, in order to further simplify the transmission model, the "truck-route" representation is adopted as described in Short et al. (2011) and Krishnan et al. (2016). The transmission network is represented similarly to pipelines, assuming that the flow in each line can be determined by an energy balance between nodes. This approximation ignores Kirchhoff's voltage law, which dictates that the power will flow along the path of least impedance. It is also assumed that the transmission lines have a maximum capacity, and that transmission expansion is not considered. Additionally, the transmission losses are characterized by a fraction loss per mile, and are not endogeneously calculated. It is important to acknowledge, however, that the transmission infrastructure affects both location and type of generation investment. Therefore, not including transmission expansion and disregarding Kirchhoff's Voltage Law could distort the planning results (Munoz et al., 2013).

5.3 MILP Formulation

This section presents a deterministic MILP formulation organized into 4 groups of constraints: operational, investment-related, generator balances, and storage constraints. Note that if an index appears in a summation or next to a $\forall$ symbol without a corresponding set, all elements in that set are assumed.

5.3.1 Operational constraints

The energy balance (5.1) ensures that, in each sub-period s of representative day d in year t, the sum of instantaneous power $p_{i,r,t,d,s}$ generated by generator clusters i in region r plus the difference between the power flow going from regions r' to region r, $p^{\text{flow}}_{r',r,t,d,s}$, and the power flowing from region r to regions r', $p^{\text{flow}}_{r,r',t,d,s}$, plus the power discharged from all the storage clusters j in region r, $p^{\text{discharge}}_{j,r,t,d,s}$, equals the load demand $L_{r,t,d,s}$ at that region r, plus the power being charged to the storage clusters j in region r, $p^{\text{charge}}_{j,r,t,d,s}$, plus a slack for curtailment of renewable generation $cu_{r,t,d,s}$.

$$\begin{aligned}\sum_{i}(p_{i,r,t,d,s}) + \sum_{r' \neq r}\left(p^{\text{flow}}_{r',r,t,d,s} \cdot (1 - T^{\text{loss}}_{r,r'} \cdot D_{r,r'}) - p^{flow}_{r,r',t,d,s}\right) + \sum_{j} p^{\text{discharge}}_{j,r,t,d,s} \\ = L_{r,t,d,s} + \sum_{j} p^{\text{charge}}_{j,r,t,d,s} + cu_{r,t,d,s} \quad \forall\ r,t,d,s\end{aligned} \tag{5.1}$$

The distance between regions $D_{r,r'}$ assumes the midpoint for each region, and the transmission loss $T^{\text{loss}}_{r,r'}$ is approximated by a fraction loss per mile.

The capacity factor constraint (5.2) limits the power outlet $p_{i,r,t,d,s}$ of renewable generators to be equal to a fraction $Cf_{i,r,t,d,s}$ of the nameplate capacity $Qg^{\text{np}}_{i,r}$ in each sub-period s of representative day d in year t, where $ngo^{\text{rn}}_{i,r,t}$ represents the number of renewable generators that are operational in year t. Due to the flexibility in sizes for renewable generators, $ngo^{\text{rn}}_{i,r,t}$ is relaxed to be continuous.

$$p_{i,r,t,d,s} = Qg^{\text{np}}_{i,r} \cdot Cf_{i,r,t,d,s} \cdot ngo^{\text{rn}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{RN}}_{r}, r, t, d, s \tag{5.2}$$

The unit commitment constraint (5.3) computes the number of generators that are ON, $u_{i,r,t,d,s}$, or in startup, $su_{i,r,t,d,s}$, and shutdown, $sd_{i,r,t,d,s}$, modes in cluster i in sub-period s of representative day d of year t, and treated as integer variables.

$$u_{i,r,t,d,s} = u_{i,r,t,d,s-1} + su_{i,r,t,d,s} - sd_{i,r,t,d,s} \qquad \forall\ i \in \mathcal{I}_r^{\mathrm{TH}}, r, t, d, s \tag{5.3}$$

The ramping limit constraints (5.4)-(5.5) capture the limitation on how fast thermal units can adjust their output power, $p_{i,r,t,d,s}$, where $Ru_i^{\max}$ is the maximum ramp-up rate, $Rd_i^{\max}$ is the maximum ramp-down rate, and $Pg_i^{\min}$ is the minimum operating limit (Palmintier and Webster, 2014).

$$\begin{aligned} p_{i,r,t,d,s} - p_{i,r,t,d,s-1} \leq \quad & Ru_i^{\max} \cdot Hs \cdot Qg_{i,r}^{\mathrm{np}} \cdot (u_{i,r,t,d,s} - su_{i,r,t,d,s}) \\ & + \max\left(Pg_i^{\min}, Ru_i^{\max} \cdot Hs\right) \cdot Qg_{i,r}^{\mathrm{np}} \cdot su_{i,r,t,d,s} \quad \forall\ i \in \mathcal{I}_r^{\mathrm{TH}}, r, t, d, s \end{aligned} \tag{5.4}$$

$$\begin{aligned} p_{i,r,t,d,s-1} - p_{i,r,t,d,s} \leq \quad & Rd_i^{\max} \cdot Hs \cdot Qg_{i,r}^{\mathrm{np}} \cdot (u_{i,r,t,d,s} - su_{i,r,t,d,s}) \\ & + \max\left(Pg_i^{\min}, Rd_i^{\max} \cdot Hs\right) \cdot Qg_{i,r}^{\mathrm{np}} \cdot sd_{i,r,t,d,s} \quad \forall\ i \in \mathcal{I}_r^{\mathrm{TH}}, r, t, d, s \end{aligned} \tag{5.5}$$

Note that the first terms on the right hand side of (5.4) and (5.5) apply only for normal operating mode (i.e., generator is ON), while the second terms apply for the startup and shutdown modes. This means that generators in normal operating mode have their ramp rates limited by $Ru_i^{\max}$ and $Rd_i^{\max}$, while generators in startup and shutdown modes have their ramp rates limited by the least restrictive between $Pg_i^{\min}$ and $Ru_i^{\max}$, $Rd_i^{\max}$ such that their operating limits (Equations 5.6 and 5.7) are still satisfied.

The operating limits constraints (5.6)-(5.7) specify that each thermal generator is either OFF and outputting zero power, or ON and running within the operating limits $Pg_i^{\min} \cdot Qg_{i,r}^{\mathrm{np}}$ and $Qg_{i,r}^{\mathrm{np}}$. The variable $u_{i,r,t,d,s}$ (integer variable) represents the number of generators that are ON in cluster $i \in \mathcal{I}_r^{\mathrm{TH}}$ at the time period t, representative day d, and sub-period s.

$$u_{i,r,t,d,s} \cdot Pg_i^{\min} \cdot Qg_{i,r}^{\mathrm{np}} \leq p_{i,r,t,d,s} \qquad \forall\ i \in \mathcal{I}_r^{\mathrm{TH}}, r, t, d, s \tag{5.6}$$

$$p_{i,r,t,d,s} + q_{i,r,t,d,s}^{\mathrm{spin}} \leq u_{i,r,t,d,s} \cdot Qg_{i,r}^{\mathrm{np}} \qquad \forall\ i \in \mathcal{I}_r^{\mathrm{TH}}, r, t, d, s \tag{5.7}$$

The upper limit constraint is modified in order to capture the need for generators to run below the maximum considering operating reserves, where $q^{\text{spin}}_{i,r,t,d,s}$ is a variable representing the spinning reserve capacity.

The total operating reserve constraint (5.8) dictates that the total spinning reserve, $q^{\text{spin}}_{i,r,t,d,s}$, plus quick-start reserve, $q^{\text{Qstart}}_{i,r,t,d,s}$, must exceed the minimum operating reserve, $Op^{\min}$, which is a percentage of the load $L_{r,t,d,s}$ in a reserve sharing region r at each sub-period s.

$$\sum_{i \in \mathcal{I}^{\text{TH}}_r} \left(q^{\text{spin}}_{i,r,t,d,s} + q^{\text{Qstart}}_{i,r,t,d,s} \right) \geq Op^{\min} \cdot L_{r,t,d,s} \qquad \forall\ r, t, d, s \tag{5.8}$$

Spinning Reserve is the on-line reserve capacity that is synchronized to the grid system and ready to meet electric demand within 10 minutes of a dispatch instruction by the independent system operator (ISO). Quick-start (or non-spinning) reserve is the extra generation capacity that is not currently connected to the system but can be brought on-line after a short delay.

The total spinning reserve constraint (5.9) specifies that the total spinning reserve $q^{\text{spin}}_{i,r,t,d,s}$ must exceed the minimum spinning reserve, $Spin^{\min}$, which is a percentage of the load $L_{r,t,d,s}$ in a reserve sharing region r at each sub-period s.

$$\sum_{i \in \mathcal{I}^{\text{TH}}_r} q^{\text{spin}}_{i,r,t,d,s} \geq Spin^{\min} \cdot L_{r,t,d,s} \qquad \forall\ r, t, d, s \tag{5.9}$$

Our model does not currently impose a minimum requirement for total quick-start reserve, as presented in Flores-Quiroz et al. (2016). However, this constraint could be easily incorporated in the formulation to address the extra secondary (quick-start) reserve requirements needed to account for the increasing short term uncertainty due to more renewable generators contributing to the grid.

The maximum spinning reserve constraint (5.10) states that the maximum fraction of capacity of each generator cluster that can contribute to spinning reserves is given by $Frac^{\text{spin}}_i$, which is a fraction of the nameplate capacity $Qg_{i,r}^{\text{np}}$.

$$q_{i,r,t,d,s}^{\text{spin}} \leq u_{i,r,t,d,s} \cdot Qg_{i,r}^{\text{np}} \cdot Frac_i^{\text{spin}} \qquad \forall\, i \in \mathcal{I}_r^{\text{TH}}, r, t, d, s \tag{5.10}$$

The maximum quick-start reserve constraint dictates that the maximum fraction of the capacity of each generator cluster that can contribute to quick-start reserves is given by $Frac_i^{\text{Qstart}}$ (fraction of the nameplate capacity $Qg_{i,r}^{\text{np}}$), and that quick-start reserves can only be provided by the generators that are OFF, i.e., not active.

$$q_{i,r,t,d,s}^{\text{Qstart}} \leq (ngo_{i,r,t}^{\text{th}} - u_{i,r,t,d,s}) \cdot Qg_{i,r}^{\text{np}} \cdot Frac_i^{\text{Qstart}} \qquad \forall\, i \in \mathcal{I}_r^{\text{TH}}, r, t, d, s \tag{5.11}$$

Here the integer variable $ngo_{i,r,t}^{\text{th}}$ represents the number of thermal generators that are operational (i.e., installed and ready to operate) at year t.

5.3.2 Investment-related constraints

The planning reserve requirement (5.12) ensures that the operating capacity is greater than or equal to the annual peak load $L_t^{\max}$, plus a predefined fraction of reserve margin $R_t^{\min}$ of the annual peak load $L_t^{\max}$.

$$\sum_{i\in\mathcal{I}_r^{\text{RN}}} \sum_{r} \left(Qg_{i,r}^{\text{np}} \cdot Q_i^{\text{v}} \cdot ngo_{i,r,t}^{\text{rn}}\right) + \sum_{i\in\mathcal{I}_r^{\text{TH}}} \sum_{r} \left(Qg_{i,r}^{\text{np}} \cdot ngo_{i,r,t}^{\text{th}}\right) \geq (1 + R_t^{\min}) \cdot L_t^{\max} \qquad \forall\, t \tag{5.12}$$

For all thermal generators, their full nameplate capacity $Qg_{i,r}^{\text{np}}$ counts towards the planning reserve requirement. However, for the renewable technologies (wind, PV and CSP), their contribution is less than the nameplate due to the inability to control dispatch and the uncertainty of the output (Short et al., 2011). Therefore, the fraction of the capacity that can be reliably counted towards the planning reserve requirement is referred to as the capacity value Q_i^{v}.

The minimum annual renewable generation requirement (5.13) ensures that, in case of policy mandates, the renewable generation quota target, $RN_t^{\min}$, which is a fraction of the energy demand ED_t, is satisfied. If not, i.e, if there is a deficit def_t^{rn} from the quota, this is subjected to a penalty that is included later in the objective function.

$$\sum_d \sum_s \left[W_d \cdot Hs \cdot \left(\sum_{i \in \mathcal{I}_r^{\text{RN}}} \sum_r p_{i,r,t,d,s} - cu_{r,t,d,s} \right) \right] + def_t^{\text{rn}} \geq RN_t^{\min} \cdot ED_t \quad \forall\, t \tag{5.13}$$

Here W_d represents the weight of the representative day d, Hs is the length of the sub-period, $cu_{r,t,d,s}$ is the curtailment of renewable generation, and ED_t represent the energy demand in year t:

$$ED_t = \sum_r \sum_d \sum_s (W_d \cdot Hs \cdot L_{r,t,d,s})$$

The maximum yearly installation constraints (5.14)-(5.15) limit the yearly installation per generation type in each region r to an upper bound $Q_{i,t}^{\text{inst,UB}}$ in MW/year. Here $ngb_{i,r,t}^{\text{rn}}$ and $ngb_{i,r,t}^{\text{th}}$ represent the number of renewable and thermal generators built in region r in year t, respectively. Note that due to the flexibility in sizes for renewable generators, $ngb_{i,r,t}^{\text{rn}}$ is relaxed to be continuous.

$$\sum_r ngb_{i,r,t}^{\text{rn}} \leq Q_{i,t}^{\text{inst,UB}} / Qg_{i,r}^{\text{np}} \qquad \forall\, i \in \mathcal{I}_r^{\text{Rnew}}, t \tag{5.14}$$

$$\sum_r ngb_{i,r,t}^{\text{th}} \leq Q_{i,t}^{\text{inst,UB}} / Qg_{i,r}^{\text{np}} \qquad \forall\, i \in \mathcal{I}_r^{\text{Tnew}}, t \tag{5.15}$$

5.3.3 Generator balance constraints

Concerning renewable generator clusters, we define a set of constraints (5.16)-(5.17) to compute the number of generators in cluster i that are ready to operate $ngo_{i,r,t}^{\text{rn}}$, taking into account the generators that were already existing at the beginning of the planning horizon $Ng_{i,r}^{\text{Rold}}$, the generators built $ngb_{i,r,t}^{\text{rn}}$, and the generators retired $ngr_{i,r,t}^{\text{rn}}$ at year t. It is important to highlight that we assume *no lead time* between the decision to build/install a generator and the moment it can begin producing electricity.

$$ngo^{\text{rn}}_{i,r,t} = Ng^{\text{Rold}}_{i,r} + ngb^{\text{rn}}_{i,r,t} - ngr^{\text{rn}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{RN}}_r, r, t = 1 \tag{5.16}$$

$$ngo^{\text{rn}}_{i,r,t} = ngo^{\text{rn}}_{i,r,t-1} + ngb^{\text{rn}}_{i,r,t} - ngr^{\text{rn}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{RN}}_r, r, t > 1 \tag{5.17}$$

As aforementioned, due to the flexibility in sizes for renewable generators, $ngo^{\text{rn}}_{i,r,t}$, $ngb^{\text{rn}}_{i,r,t}$, and $ngr^{\text{rn}}_{i,r,t}$ are relaxed to be continuous. Note that $ngb^{\text{rn}}_{i,r,t}$ for $i \in \mathcal{I}^{\text{Rold}}_r$ is fixed to zero in all time periods, i.e., the clusters of existing renewable generators cannot have any new additions during the time horizon considered.

We also define a set of constraints (5.18)-(5.19) to enforce the renewable generators that reached the end of their lifetime to either retire, $ngr^{\text{rn}}_{i,r,t}$, or have their life extended, $nge^{\text{rn}}_{i,r,t}$. $Ng^{\text{r}}_{i,r,t}$ is a parameter that represents the number of old generators (i.e., $i \in \mathcal{I}^{\text{old}}_r$) that reached the end of their lifetime, LT_i, at year t.

$$Ng^{\text{r}}_{i,r,t} = ngr^{\text{rn}}_{i,r,t} + nge^{\text{rn}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{Rold}}_r, r, t \tag{5.18}$$

$$\sum_{t'' \leq t - LT_i} ngb^{\text{rn}}_{i,r,t''} = \sum_{t' \leq t} \left(ngr^{\text{rn}}_{i,r,t'} + nge^{\text{rn}}_{i,r,t'} \right) \qquad \forall\ i \in \mathcal{I}^{\text{Rnew}}_r, r, t \tag{5.19}$$

Concerning thermal generator clusters, we define a set of constraints (5.20)-(5.21) to compute the number of generators in cluster i that are ready to operate $ngo^{\text{th}}_{i,r,t}$, taking into account the generators that were already existing at the beginning of the planning horizon $Ng^{\text{Told}}_{i,r}$, the generators built $ngb^{\text{th}}_{i,r,t}$, and the generators retired $ngr^{\text{th}}_{i,r,t}$ at year t.

$$ngo^{\text{th}}_{i,r,t} = Ng^{\text{Told}}_{i,r} + ngb^{\text{th}}_{i,r,t} - ngr^{\text{th}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{TH}}_r, r, t = 1 \tag{5.20}$$

$$ngo^{\text{th}}_{i,r,t} = ngo^{\text{th}}_{i,r,t-1} + ngb^{\text{th}}_{i,r,t} - ngr^{\text{th}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{TH}}_r, r, t > 1 \tag{5.21}$$

Note that $ngb^{\text{th}}_{i,r,t}$ for $i \in \mathcal{I}^{\text{Told}}_r$ is fixed to zero in all time periods, i.e., the clusters of existing thermal generators cannot have any new additions during the time horizon considered.

We also define a set of constraints (5.22)-(5.23) to enforce the thermal generators that reached the end of their lifetime to either retire, $ngr^{\text{th}}_{i,r,t}$, or have their life extended $nge^{\text{th}}_{i,r,t}$.

$$Ng^{\text{r}}_{i,r,t} = ngr^{\text{th}}_{i,r,t} + nge^{\text{th}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{Told}}_r, r, t \tag{5.22}$$

$$\sum_{t'' \leq t - LT_i} ngb^{\text{th}}_{i,r,t''} = \sum_{t' \leq t} \left(ngr^{\text{th}}_{i,r,t'} + nge^{\text{th}}_{i,r,t'}\right) \qquad \forall\ i \in \mathcal{I}^{\text{Tnew}}_r, r, t \tag{5.23}$$

Finally, we have constraint (5.24) that ensures that only installed generators can be in operation:

$$u_{i,r,t,d,s} \leq ngo^{\text{th}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{Tnew}}_r, r, t, d, s \tag{5.24}$$

5.3.4 Storage constraints

We also include a set constraints related to the energy storage devices, which are assumed to be ideal and generic (Pozo et al., 2014). Constraints (5.25)-(5.26) compute the number of storage units that are ready to operate $nso_{j,r,t}$, taking into account the storage units already existing at the beginning of the planning horizon $Ns_{j,r}$ and the ones built $nsb_{j,r,t}$ and retired $nsr_{j,r,t}$ at year t. Due to the flexibility in sizes for storage units, $nso_{j,r,t}$, $nsb_{j,r,t}$, and $nsr_{j,r,t}$ are relaxed to be continuous.

$$nso_{j,r,t} = Ns_{j,r} + nsb_{j,r,t} - nsr_{s,r,t} \qquad \forall\ j, r, t = 1 \tag{5.25}$$

$$nso_{j,r,t} = nso_{j,r,t-1} + nsb_{j,r,t} - nsr_{j,r,t} \qquad \forall\ j, r, t > 1 \tag{5.26}$$

Constraint (5.27) enforces retirement of storage units that have reached the end of their lifetime,LT^{s}_j.

$$\sum_{t'' \leq t - LT^{\text{s}}_j} nsb_{j,r,t''} = \sum_{t' \leq t} nsr_{j,r,t'} \qquad \forall\ j, r, t \tag{5.27}$$

Constraints (5.28) and (5.29) establish that the power charge, $p^{\text{charge}}_{j,r,t,d,s}$, and discharge, $p^{\text{discharge}}_{j,r,t,d,s}$, of the storage units in cluster j, $nso_{j,r,t}$, has to be within the operating limits:

$Charge_j^{\min}$ and $Charge_j^{\max}$, and $Discharge_j^{\min}$ and $Discharge_j^{\min}$, respectively.

$$Charge_j^{\min} \cdot nso_{j,r,t} \leq p_{j,r,t,d,s}^{\text{charge}} \leq Charge_j^{\max} \cdot nso_{j,r,t} \qquad \forall\ j,r,t,d,s \tag{5.28}$$

$$Discharge_j^{\min} \cdot nso_{j,r,t} \leq p_{j,r,t,d,s}^{\text{discharge}} \leq Discharge_j^{\max} \cdot nso_{j,r,t} \qquad \forall\ j,r,t,d,s \tag{5.29}$$

Constraint (5.30) specifies that the energy storage level, $p_{j,r,t,d,s}^{\text{level}}$, for the storage units in cluster j, $nso_{j,r,t}$ has to be within the storage capacity limits $Storage_j^{\min}$ and $Storage_j^{\max}$.

$$Storage_j^{\min} \cdot nso_{j,r,t} \leq p_{j,r,t,d,s}^{\text{level}} \leq Storage_j^{\max} \cdot nso_{j,r,t} \qquad \forall\ j,r,t,d,s \tag{5.30}$$

Constraints (5.31) and (5.32) show the power balance in the storage units. The state of charge $p_{j,r,t,d,s}^{\text{level}}$ at the end of sub-period s depends on the previous state of charge $p_{j,r,t,d,s-1}^{\text{level}}$, and the power charged $p_{j,r,t,d,s}^{\text{charge}}$ and discharged $p_{j,r,t,d,s}^{\text{discharge}}$ at sub-period s. The symbols η_j^{charge} and $\eta_j^{\text{discharge}}$ represent the charging and discharging efficiencies, respectively. For the first hour of the day d of year t, the previous state of charge (i.e., $s = 0$) is the variable $p_{j,r,t,d}^{\text{level},0}$.

$$p_{j,r,t,d,s}^{\text{level}} = p_{j,r,t,d,s-1}^{\text{level}} + \eta_j^{\text{charge}} \cdot p_{j,r,t,d,s}^{\text{charge}} + p_{j,r,t,d,s}^{\text{discharge}}/\eta_j^{\text{discharge}} \qquad \forall\ j,r,t,d,s>1 \tag{5.31}$$

$$p_{j,r,t,d,s}^{\text{level}} = p_{j,r,t,d}^{\text{level},0} + \eta_j^{\text{charge}} \cdot p_{j,r,t,d,s}^{\text{charge}} + p_{j,r,t,d,s}^{\text{discharge}}/\eta_j^{\text{discharge}} \qquad \forall\ j,r,t,d,s=1 \tag{5.32}$$

Constraints (5.33) and (5.34) force the storage units to begin $p_{j,r,t,d}^{\text{level},0}$ and end $p_{j,r,t,d,s=S}^{\text{level}}$ each day d of year t with 50% of their maximum storage $Storage_j^{\max}$. This is a heuristic to attach carryover storage level form one representative day to the next (Liu et al., 2017).

$$p_{j,r,t,d}^{\text{level},0} = 0.5 \cdot Storage_j^{\max} \cdot nso_{j,r,t} \qquad \forall\ j,r,t,d \tag{5.33}$$

$$p_{j,r,t,d,s}^{\text{level}} = 0.5 \cdot Storage_j^{\max} \cdot nso_{j,r,t} \qquad \forall\ j,r,t,d,s=S \tag{5.34}$$

5.3.5 Objective Function

The objective of this model is to minimize the net present cost, Φ, over the planning horizon, which includes operating costs Φ^{opex}, investment costs Φ^{capex}, and potential penalties

Φ^{PEN} for not meeting the the targets on renewables.

$$\min \quad \Phi = \sum_t \left(\Phi_t^{\text{opex}} + \Phi_t^{\text{capex}} + \Phi_t^{\text{PEN}}\right) \tag{5.35}$$

The operating expenditure, Φ_t^{opex}, comprises the variable $VOC_{i,t}$ and fixed $FOC_{i,t}$ operating costs, as well as fuel cost P_i^{fuel} per heat rate HR_i, carbon tax $Tx_t^{\text{CO}_2}$ for CO_2 emissions $EF_i^{CO_2}$, and start-up cost (variable cost P_i^{fuel} that depends on the amount of fuel burned for startup F_i^{start}, and fixed cost C_i^{start}).

$$\begin{aligned}
\Phi_t^{\text{opex}} = \quad & If_t \cdot \Bigg[\sum_d \sum_s W_d \cdot hs \cdot \\
& \left(\sum_i \sum_r (VOC_{i,t} + P_i^{\text{fuel}} \cdot HR_i + Tx_t^{\text{CO}_2} \cdot EF_i^{CO_2} \cdot HR_i) \cdot p_{i,r,t,d,s} \right) \\
& + \left(\sum_{i \in \mathcal{I}_r^{\text{RN}}} \sum_r FOC_{i,t} \cdot Qg_{i,r}^{\text{np}} \cdot ngo_{i,r,t}^{\text{rn}} \right) \\
& + \left(\sum_{i \in \mathcal{I}_r^{\text{TH}}} \sum_r FOC_{i,t} \cdot Qg_{i,r}^{\text{np}} \cdot ngo_{i,r,t}^{\text{th}} \right) \\
& + \sum_{i \in \mathcal{I}_r^{\text{TH}}} \sum_r \sum_d \sum_s W_d \cdot Hs \cdot su_{i,r,t,d,s} \cdot Qg_{i,r}^{\text{np}} \\
& \cdot \left(F_i^{\text{start}} \cdot P_i^{\text{fuel}} + F_i^{\text{start}} \cdot EF^{CO_2} \cdot Tx_t^{\text{CO}_2} + C_i^{\text{start}}\right) \Bigg]
\end{aligned} \tag{5.36}$$

The capital expenditure, Φ_t^{capex}, includes the amortized cost of acquiring new generators, $DIC_{i,t}$, new storage devices, $SIC_{j,t}$, and the amortized cost of extending the life of generators that reached their expected lifetime. The latter is assumed to be a fraction LE_i of the investment cost, $DIC_{i,t}$, in a new generator with the same or equivalent generation technology. In this framework, the investment cost takes into account the remaining value at the end of the time horizon by considering the annualized capital cost and multiplying it by the number of years remaining in the planning horizon at the time of installation to calculate the $DIC_{i,t}$.

$$
\begin{aligned}
\Phi_t^{\text{capex}} = \quad & If_t \cdot \Bigg[\sum_{i \in \mathcal{I}_r^{\text{Rnew}}} \sum_r DIC_{i,t} \cdot CC_i^{\text{m}} \cdot Qg_{i,r}^{\text{np}} \cdot ngb_{i,r,t}^{\text{rn}} \\
& + \sum_{i \in \mathcal{I}_r^{\text{Tnew}}} \sum_r DIC_{i,t} \cdot CC_i^{\text{m}} \cdot Qg_{i,r}^{\text{np}} \cdot ngb_{i,r,t}^{\text{th}} \\
& + \sum_j \sum_r SIC_{j,t} \cdot Storage_j^{\text{max}} \cdot nsb_{j,r,t} \\
& + \sum_{i \in \mathcal{I}_r^{\text{RN}}} \sum_r DIC_{i,t} \cdot LE_i \cdot Qg_{i,r}^{\text{np}} \cdot nge_{i,r,t}^{\text{rn}} \\
& + \sum_{i \in \mathcal{I}_r^{\text{TH}}} \sum_r DIC_{i,t} \cdot LE_i \cdot Qg_{i,r}^{\text{np}} \cdot nge_{i,r,t}^{\text{th}} \Bigg]
\end{aligned}
\tag{5.37}
$$

The capital multiplier CC_i^{m} associated with new generator clusters is meant to account for differences in depreciation schedules applicable to each technology, with higher values being indicative of slower depreciating schedule and vice versa.

Lastly, the penalty cost, Φ_t^{PEN}, includes the potential fines for not meeting the renewable energy quota, PEN_t^{rn}, and curtailing the renewable generation.

$$
\Phi_t^{\text{PEN}} = \quad If_t \cdot \left(PEN_t^{\text{rn}} \cdot def_t^{\text{rn}} + PEN^c \cdot \sum_r \sum_d \sum_s cu_{r,t,d,s} \right) \tag{5.38}
$$

The parameters If_t, $DIC_{i,t}$, $ACC_{i,t}$, and T_t^{remain} are defined in B.1.

The integrated planning and operations model for the electric power systems is then given by the multi-period MILP model defined by equations (5.1)-(5.38).

5.4 Nested Decomposition for Multiperiod MILP Problems

Even though the multi-period MILP formulation in Section 5.3 incorporates modeling strategies to reduce the size of the model, it can still be very expensive to solve and potentially intractable depending on the size of the area considered, and the time resolution of the representative periods per season. Therefore, we propose a decomposition algorithm based on Nested Benders Decomposition (Birge, 1985), Stochastic Dual Dynamic Programming (SDDP) (Pereira and Pinto, 1991), and Generalized Dual Dynamic Programming (GDDP) (Velásquez Bermúdez, 2002).

These methods are used in the context of Multistage Linear Stochastic Programming (MLSP), but their major limitation is that they can only be applied to convex subproblems. Thus, they are not suitable for our problem, which gives rise to mixed-integer subproblems. In this context, Cerisola et al. (2009) propose a variant of Benders Decomposition for multistage stochastic integer programming and apply it to the stochastic unit commitment problem. Thome et al. (2013) introduce an extension of the SDDP framework by using Lagrangean Relaxation to convexify the recourse function applied to nonconvex hydrothermal operation planning. Zou et al. (2018b) present a valid Stochastic Dual Dynamic Integer Programming (SDDiP) algorithm for Multistage Stochastic Integer Programming (MSIP) with binary state variables, and prove that for some of the cuts presented the algorithm converges in a finite number of steps. In their more recent paper, Zou et al. (2017) apply their SDDiP algorithm to Stochastic Unit Commitment problems. Finally, Steeger and Rebennack (2017) propose a dynamic convexification within Benders Decomposition using Lagrangean relaxation and apply it to the electricity market bidding problem.

In this work, we have adapted the algorithm proposed by Zou et al. (2018b) for deterministic multi-period MILP models and apply it to the formulation given in Section 5.3. Moreover, we modified the approach in Zou et al. (2018b) by allowing for integer and continuous state variables, at the expense of losing its finite convergence property due to potential duality gap. The novel features of the proposed decomposition are: a) targets multi-period deterministic problems; b) allows integer and continuous state variables; c) allows links between non-adjacent time-periods; and d) utilizes an acceleration technique to generate warm-start cuts.

To facilitate the understanding of the algorithm, we first describe how the multi-period

MILP model defined by equations (5.1)-(5.38) is decomposed by time period (year). Then, we introduce a more concise notation to represent the decomposed 1-year-long MILPs, and use this notation to describe the Nested Decomposition algorithm in Section 5.4.2.

5.4.1 Decomposition by time period (year)

In our formulation, the only constraints that depend on more than one time period are equations (5.17), (5.19), (5.21), (5.23), (5.26), and (5.27). Therefore, these constraints have to be reformulated in order to be able to solve the problem separately for each time period, which is done by duplicating the linking variables, $ngo^{\text{rn}}_{i,r,t}$, $ngo^{\text{th}}_{i,r,t}$, $ngb^{\text{rn}}_{i,r,t}$, $ngb^{\text{th}}_{i,r,t}$, $nso_{j,r,t}$, and $nsb_{j,r,t}$.

Equation (5.17), which computes the number of renewable generators that are operational at time period t based on the number of generators built and retired at t and operational at $t-1$, is substituted by equations (5.39) and (5.40).

$$ngo^{\text{rn}}_{i,r,t} = ngo^{\text{rn,prev}}_{i,r,t} + ngb^{\text{rn}}_{i,r,t} - ngr^{\text{rn}}_{i,r,t} \qquad \forall\, i \in \mathcal{I}^{\text{RN}}_r, r, t > 1 \tag{5.39}$$

$$ngo^{\text{rn,prev}}_{i,r,t} = n\hat{g}o^{\text{rn}}_{i,r,t-1} \qquad \leftarrow \mu^{\text{o,rn}}_{i,r,t} \in \mathbb{R}^{|\mathcal{I}^{\text{RN}}_r|+|\mathcal{R}|+|\mathcal{T}|-1} \qquad \forall\, i \in \mathcal{I}^{\text{RN}}_r, r, t > 1 \tag{5.40}$$

Here $ngo^{\text{rn,prev}}_{i,r,t}$ is the duplicated variable representing $ngo^{\text{rn}}_{i,r,t-1}$, and $n\hat{g}o^{\text{rn}}_{i,r,t-1}$ is the solution for $ngo^{\text{rn}}_{i,r,t}$ at time period $t-1$, which is fixed when solving time period t. The Lagrange multiplier $\mu^{\text{o,rn}}_{i,r,t}$ of equation (5.40) is unrestricted in sign.

Similarly, equation (5.21), which refers to thermal generators, is substituted by equations (5.41) and (5.42).

$$ngo^{\text{th}}_{i,r,t} = ngo^{\text{th,prev}}_{i,r,t} + ngb^{\text{th}}_{i,r,t} - ngr^{\text{th}}_{i,r,t} \qquad \forall\, i \in \mathcal{I}^{\text{TH}}_r, r, t > 1 \tag{5.41}$$

$$ngo^{\text{th,prev}}_{i,r,t} = n\hat{g}o^{\text{th}}_{i,r,t-1} \qquad \leftarrow \mu^{\text{o,th}}_{i,r,t} \in \mathbb{R}^{|\mathcal{I}^{\text{TH}}_r|+|\mathcal{R}|+|\mathcal{T}|-1} \qquad \forall\, i \in \mathcal{I}^{\text{TH}}_r, r, t > 1 \tag{5.42}$$

Here $ngo^{\text{th,prev}}_{i,r,t}$ is the duplicated variable representing $ngo^{\text{th}}_{i,r,t-1}$, and $n\hat{g}o^{\text{th}}_{i,r,t-1}$ is the solution for $ngo^{\text{th}}_{i,r,t}$ at time period $t-1$, which is fixed when solving time period t. The Lagrange multiplier $\mu^{\text{o,th}}_{i,r,t}$ of equation (5.42) is unrestricted in sign.

Analogously, equation (5.26) that refers to storage units is substituted by equations (5.43) and (5.44).

$$nso_{i,r,t} = nso^{\text{prev}}_{j,r,t} + nsb_{j,r,t} - nsr_{j,r,t} \qquad \forall\ j,r,t>1 \tag{5.43}$$

$$nso^{\text{prev}}_{j,r,t} = \hat{nso}_{j,r,t-1} \qquad \leftarrow \mu^{\text{o,s}}_{j,r,t} \in \mathbb{R}^{|\mathcal{J}|+|\mathcal{R}|+|\mathcal{T}|-1} \qquad \forall\ j,r,t>1 \tag{5.44}$$

Here $nso^{\text{prev}}_{j,r,t}$ is the duplicated variable representing $nso_{i,r,t-1}$, and $\hat{nso}_{j,r,t-1}$ is the solution for $nso_{i,r,t}$ at time period $t-1$, which is fixed when solving time period t. The Lagrange multiplier $\mu^{\text{o,s}}_{j,r,t}$ of equation (5.44) is unrestricted in sign.

Equations (5.19) and (5.23) compute the age of the new generators that are built during the planning horizon to be able to enforce their retirement (or life extension) when they achieve the end of their lifetime. Hence, these constraints link time period t to time period $t - LT_i$, where LT_i is the expected lifetime of a generator in cluster i. In order to decouple those time periods, (5.19) has to be replaced by (5.45) and (5.46).

$$ngb^{\text{rn,LT}}_{i,r,t} = ngr^{\text{rn}}_{i,r,t} + nge^{\text{rn}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{Rnew}}_r, r, t \tag{5.45}$$

$$ngb^{\text{rn,LT}}_{i,r,t} = \hat{ngb}^{\text{rn}}_{i,r,t-LT_i} \qquad \leftarrow \mu^{\text{b,rn}}_{i,r,t} \in \mathbb{R}^{|\mathcal{I}^{\text{Rnew}}_r|+|\mathcal{R}|+|\mathcal{T}|} \qquad \forall\ i \in \mathcal{I}^{\text{Rnew}}_r, r, t \tag{5.46}$$

Here $ngb^{\text{rn,LT}}_{i,r,t}$ is the duplicated variable representing the renewable generators built in year $t-LT_i$, $ngb^{\text{rn}}_{i,r,t-LT_i}$. Thus, the model is able to track when the end of the generators' lifetime is, which is the year they were built plus their lifetime, LT_i. The solution for $ngb^{\text{rn}}_{i,r,t}$ at year $t-LT_i$, $\hat{ngb}^{\text{rn}}_{i,r,t-LT_i}$, is fixed when solving time period t. The Lagrange multiplier $\mu^{\text{b,rn}}_{i,r,t}$ of equation (5.46) is unrestricted in sign.

Similarly, constraint (5.23) is replaced by (5.47) and (5.48).

$$ngb^{\text{th,LT}}_{i,r,t} = ngr^{\text{th}}_{i,r,t} + nge^{\text{th}}_{i,r,t} \qquad \forall\ i \in \mathcal{I}^{\text{Tnew}}_r, r, t \tag{5.47}$$

$$ngb^{\text{th,LT}}_{i,r,t} = \hat{ngb}^{\text{th}}_{i,r,t-LT_i} \qquad \leftarrow \mu^{\text{b,th}}_{i,r,t} \in \mathbb{R}^{|\mathcal{I}^{\text{Tnew}}_r|+|\mathcal{R}|+|\mathcal{T}|} \qquad \forall\ i \in \mathcal{I}^{\text{Tnew}}_r, r, t \tag{5.48}$$

Here $ngb^{\text{th,LT}}_{i,r,t}$ is the duplicated variable representing the thermal generators built at year

$t - LT_i$, $ngb^{\text{th}}_{i,r,t-LT_i}$. The solution for $ngb^{\text{th}}_{i,r,t}$ at time period $t - LT_i$, $\hat{ngb}^{\text{th}}_{i,r,t-LT_i}$, is fixed when solving time period t. The Lagrange multiplier $\mu^{\text{b,th}}_{i,r,t}$ of equation (5.48) is unrestricted in sign.

Equation (5.27) computes the age of the new storage units that are built during the planning horizon to be able to enforce their retirement when they achieve the end of their lifetime. Hence, this constraint links time period t to time period $t - LT^{\text{s}}_j$, where LT^{s}_j is the expected lifetime of a storage device in cluster j. In order to decouple those time periods, (5.27) has to be replaced by (5.49) and (5.46).

$$nsb^{\text{LT}}_{j,r,t} = nsr_{j,r,t} \qquad \forall\, j,r,t \tag{5.49}$$

$$nsb^{\text{LT}}_{j,r,t} = \hat{nsb}_{i,r,t-LT^{\text{s}}_j} \qquad \leftarrow \mu^{\text{b,s}}_{j,r,t} \in \mathbb{R}^{|\mathcal{J}|+|\mathcal{R}|+|\mathcal{T}|} \qquad \forall\, j,r,t \tag{5.50}$$

Here $nsb^{\text{LT}}_{j,r,t}$ is the duplicated variable representing the storage units built in year $t - LT^{\text{s}}_j$, $nsb_{j,r,t-LT^{\text{s}}_j}$. Thus, the model is able to track when the end of the storage units' lifetime is, which is the year they were installed plus their lifetime, LT^{s}_j. The solution for $nsb_{j,r,t}$ at year $t - LT^{\text{s}}_j$, $\hat{nsb}_{j,r,t-LT^{\text{s}}_j}$, is fixed when solving time period t. The Lagrange multiplier $\mu^{\text{b,s}}_{j,r,t}$ of equation (5.50) is unrestricted in sign.

Furthermore, the objective function for a given time period is solved independently, and incorporates the cuts for future cost that are added in the following iterations. These cuts, given by equation (5.52), project the problem onto the subspace defined by the linking variables, and will be explained in detail in Section 5.4.2. Hence, equation (5.35) is replaced by (5.51)-(5.52),

$$\min\ \Phi_t = \Phi^{\text{opex}}_t + \Phi^{\text{capex}}_t + \Phi^{\text{PEN}}_t + \alpha_t \tag{5.51}$$

$$\begin{aligned}
\alpha_t \quad &\geq \hat{\Phi}_{t+1,k} + \sum_{i \in \mathcal{I}_r^{\mathrm{RN}}, r} \mu^{\mathrm{o,rn}}_{i,r,t+1,k} \cdot \left(n\hat{g}o^{\mathrm{rn}}_{i,r,t,k} - ngo^{\mathrm{rn}}_{i,r,t} \right) \\
&+ \sum_{i \in \mathcal{I}_r^{\mathrm{TH}}, r} \mu^{\mathrm{o,th}}_{i,r,t+1,k} \cdot \left(n\hat{g}o^{\mathrm{th}}_{i,r,t,k} - ngo^{\mathrm{th}}_{i,r,t} \right) \\
&+ \sum_{j,r} \mu^{\mathrm{o,s}}_{j,r,t+LT_j^{\mathrm{s}},k} \cdot \left(n\hat{s}o_{j,r,t,k} - nso_{j,r,t} \right) \\
&+ \sum_{i \in \mathcal{I}_r^{\mathrm{RN}}, r} \mu^{\mathrm{b,rn}}_{i,r,t+LT_i,k} \cdot \left(n\hat{gb}^{\mathrm{rn}}_{i,r,t,k} - ngb^{\mathrm{rn}}_{i,r,t} \right) \\
&+ \sum_{i \in \mathcal{I}_r^{\mathrm{TH}}, r} \mu^{\mathrm{b,th}}_{i,r,t+LT_i,k} \cdot \left(n\hat{gb}^{\mathrm{th}}_{i,r,t,k} - ngb^{\mathrm{th}}_{i,r,t} \right) \\
&+ \sum_{j,r} \mu^{\mathrm{b,s}}_{j,r,t+LT_j^{\mathrm{s}},k} \cdot \left(n\hat{sb}_{j,r,t,k} - nsb_{j,r,t} \right) \qquad \forall\ k
\end{aligned} \tag{5.52}$$

where k is the iteration counter.

The MILP subproblem for a given time period t and iteration k, described by equations (5.1)-(5.16), (5.18), (5.20), (5.22), (5.24), (5.25), (5.28)-(5.34), (5.36)-(5.52), can be more concisely represented by $(\mathscr{P}_{t,k})$.

$$\mathscr{P}_{t,k} : \Phi_{t,k}(\hat{x}_{t-1,k}, \phi_{t,k}) = \min_{x_t, y_t, z_t} \quad f_t(x_t, y_t) + \phi_{t,k}(\hat{x}_{t,k}) \tag{5.53a}$$

$$\text{s.t.} \quad z_t = \hat{x}_{t-1,k} \qquad \leftarrow \mu_{t,k} \in \mathbb{R}^n \tag{5.53b}$$

$$(x_t, y_t, z_t) \in \mathcal{X}_t \tag{5.53c}$$

where the feasible region $\mathcal{X}_t$ is the *mixed-integer* set given by

$$\begin{aligned}
\mathcal{X}_t = \Big\{ (x_t, y_t, z_t) : &(5.1) - (5.16), (5.18), (5.20), (5.22), (5.24), (5.25), \\
&(5.28) - (5.34), (5.36) - (5.39), (5.41), (5.43), \\
&(5.45), (5.47), (5.49) \\
&x_t \in \mathbb{Z}_+^{n_1} \times \mathbb{R}_+^{n_2}, \quad y_t \in \mathbb{Z}_+^{m_1} \times \mathbb{R}_+^{m_2}, \quad z_t \in \mathbb{R}^n \qquad \Big\}
\end{aligned} \tag{5.54}$$

and $n = n_1 + n_2$, $m = m_1 + m_2$.

The components of $(\mathscr{P}_{t,k})$ map to our problem as follows:

- x_t represents the state (i.e., linking) variables, $ngo^{\mathrm{rn}}_{i,r,t}$, $ngo^{\mathrm{th}}_{i,r,t}$, $nso_{j,r,t}$, $ngb^{\mathrm{rn}}_{i,r,t}$, $ngb^{\mathrm{th}}_{i,r,t}$, $nsb_{j,r,t}$. These are mixed integer variables since $ngo^{\mathrm{rn}}_{i,r,t}, ngb^{\mathrm{rn}}_{i,r,t}, nso_{j,r,t}, nsb_{i,r,t} \in \mathbb{R}_+$, and $ngo^{\mathrm{th}}_{i,r,t}, ngb^{\mathrm{th}}_{i,r,t} \in \mathbb{Z}_+$.
- z_t represents the duplicated state variables, $ngo^{\mathrm{rn,prev}}_{i,r,t}$, $ngo^{\mathrm{th,prev}}_{i,r,t}$, $nso^{\mathrm{prev}}_{j,r,t}$, $ngb^{\mathrm{rn,LT}}_{i,r,t}$, $ngb^{\mathrm{th,LT}}_{i,r,t}$, $nsb^{\mathrm{LT}}_{j,r,t}$, which are continuous variables.
- y_t represents the local variables, i.e. all the other variables not listed above. These are mixed-integer variables.
- $\hat{x}_{t-1,k}$ is the state of the system at the start of a time period t of iteration k, i.e., the solution for x_t obtained in a previous period of the Forward Pass that is linked to (thus fixed in) the current time period (both in iteration k). In our problem this corresponds to the fixed values of $n\hat{g}o^{\mathrm{rn}}_{i,r,t-1}$, $n\hat{g}o^{\mathrm{th}}_{i,r,t-1}$, $n\hat{s}o_{j,r,t-1}$, $\hat{ngb}^{\mathrm{rn}}_{i,r,t-LT_i}$, $\hat{ngb}^{\mathrm{th}}_{i,r,t-LT_i}$, $\hat{nsb}_{j,r,t-LT^{\mathrm{s}}_j}$ from equations (5.40), (5.42), (5.44), (5.46), (5.48), (5.50), such that $\hat{x}_{t-1,k} = (\hat{x}^{\mathrm{o}}_{t-1,k}, \hat{x}^{\mathrm{b}}_{t-LT,k})$, where $\hat{x}^{\mathrm{o}}_{t-1,k}$ maps to $(n\hat{g}o^{\mathrm{rn}}_{i,r,t-1}$, $n\hat{g}o^{\mathrm{th}}_{i,r,t-1}$, $n\hat{s}o_{j,r,t-1})$, and $\hat{x}^{\mathrm{b}}_{t-LT,k}$ maps to $(\hat{ngb}^{\mathrm{rn}}_{i,r,t-LT_i}$, $\hat{ngb}^{\mathrm{th}}_{i,r,t-LT_i}$, $\hat{nsb}_{j,r,t-LT^{\mathrm{s}}_j})$. Note that $\hat{x}^{b}_{t-LT,k}$ is a slight abuse of notation since the lifetimes LT_i and LT^{s}_j of different generator and storage clusters may differ. These fixed values are also used in the following Backward Pass.
- $f(x_t, y_t)$ is the objective function in terms of the state and local variables, x_t and y_t, respectively.
- $\phi_{t,k}$ is the cost-to-go as a function of $\hat{x}_{t,k}$.
- Constraint (5.53b) represents the equalities (5.40), (5.42), (5.44), (5.46), (5.48), (5.50), and $\mu_{t,k}$ represent their Lagrange multipliers $\mu^{\mathrm{o,rn}}_{i,r,t}$, $\mu^{\mathrm{o,th}}_{i,r,t}$, $\mu^{\mathrm{o,s}}_{j,r,t}$, $\mu^{\mathrm{b,rn}}_{i,r,t}$, $\mu^{\mathrm{b,th}}_{i,r,t}$, $\mu^{\mathrm{b,s}}_{j,r,t}$, respectively, such that $\mu_{t,k} = (\mu^{\mathrm{o}}_{t,k}, \mu^{\mathrm{b}}_{t,k})$, where $\mu^{\mathrm{o}}_{t,k}$ maps to $(\mu^{\mathrm{o,rn}}_{i,r,t}$, $\mu^{\mathrm{o,th}}_{i,r,t}$, $\mu^{\mathrm{o,s}}_{j,r,t})$, and $\mu^{\mathrm{b}}_{t,k}$ maps to $(\mu^{\mathrm{b,rn}}_{i,r,t}$, $\mu^{\mathrm{b,th}}_{i,r,t}$, $\mu^{\mathrm{b,s}}_{j,r,t})$.

The cost-to-go function, $\phi_{t,k}(\cdot)$, is defined as:

$$\phi_{t,k}(\hat{x}_{t,k}) := \min_{x_t,\alpha_t}\{\alpha_t : \ \alpha_t \geq \hat{\Phi}_{t+1,k} + \mu^{\intercal}_{t+1,k}(\hat{x}_{t,k} - x_t)\} \tag{5.55}$$

where $\hat{\Phi}_{t+1,k}$ is the optimal value for time period $t+1$, $\mu_{t+1,k}$ is the Lagrange multiplier for the equality constraint, both obtained in the Backward Pass of iteration k, such that $\mu_{t+1,k} = (\mu^{\text{o}}_{t+1,k}, \mu^{\text{b}}_{t+LT,k})$, where $\mu^{\text{o}}_{t+1,k}$ maps to $(\mu^{\text{o,rn}}_{i,r,t+1}, \mu^{\text{o,th}}_{i,r,t+1}, \mu^{\text{o,s}}_{j,r,t+1})$, and $\mu^{\text{b}}_{t+LT,k}$ maps to $(\mu^{\text{b,rn}}_{i,r,t+LT_i}, \mu^{\text{b,rn}}_{i,r,t+LT_i}, \mu^{\text{b,s}}_{j,r,t+LT^{\text{s}}_j})$. Note that $\mu^{\text{b}}_{t+LT,k}$ is a slight abuse of notation since the lifetimes LT_i and LT^{s}_j of different generator and storage clusters may differ. $\hat{x}_{t,k}$ is the solution for x_t obtained in the Forward Pass of iteration k and fixed in the following Backward Pass.

5.4.2 Description of the algorithm

The Nested Decomposition algorithm consists of decomposing the problem per time period (year) and solving it iteratively in a forward and a backward fashion. The Forward Pass yields a feasible upper bound, while the Backward Pass, which generates cuts from the relaxed subproblems, provides a lower bound. New cuts are added in the Backward Pass of each iteration k, and are kept in the following Forward Pass, until the difference between the upper and lower bounds is less than a pre-specified tolerance, ϵ_1, or iteration k has reached the maximum number of iterations, $MaxIter$, as shown in Figure 5.3. Even though this algorithm is very similar to the work by Zou et al. (2018b), we provide its full description since the mapping between a stochastic multi-stage and a deterministic multi-period formulation is not trivial.

Forward Pass

The purpose of the forward pass is to generate a feasible solution to the full problem. It accomplishes this by making decisions in time period t, implementing the investment decisions for that period, and then repeating the process in the subsequent period. Therefore, this first step consists of solving the optimization problem for each time period sequentially, using the solution from the appropriate previous time period for $\hat{x}_{t-1,k}$. This first part of the algorithm, the Forward Pass, is illustrated in Figure 5.4.

The problem is assumed to have complete continuous recourse, which means that for any value of state variable (i.e., linking variable) and local integer variables, there are values for

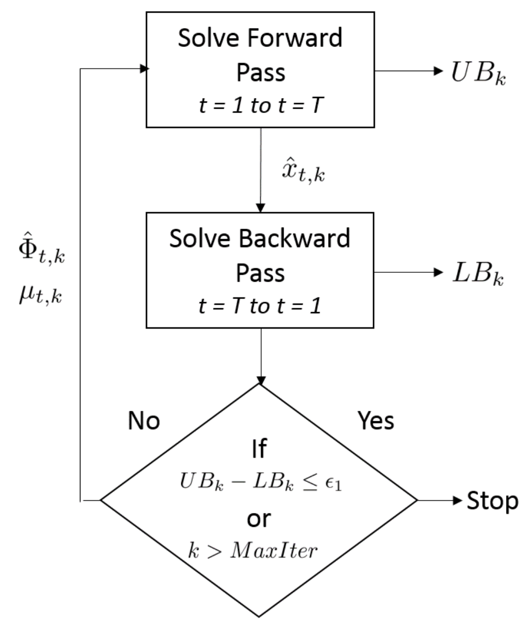

Figure 5.3: Steps at iteration k of the Nested Decomposition algorithm

the continuous local variables such that the solution is feasible. This assumption is valid since feasibility can be achieved by adding nonnegative slack variables and penalizing them in the objective function.

The upper bound, UB_k, is calculated in the Forward Pass as follows in (5.56). It is easy to see that the sum of the optimal solutions of the Forward Pass subproblems at iteration k, $\hat{\Phi}_{t,k}$, minus the cost-to-go approximations, $\hat{\alpha}_t$, for all time periods at that iteration, yields a valid upper bound, since the sequential solution of the time periods in a myopic fashion, without relaxing any constraint or integrality, gives a feasible solution to the full-space MILP problem.

$$UB_k = \sum_t \left(\hat{\Phi}_{t,k} - \hat{\alpha}_t \right) \qquad \forall\ k \tag{5.56}$$

Backward Pass

After solving the Forward subproblem for all the time periods, the next step is the Backward Pass, the purpose of which is to generate cuts. This step consists of solving the subproblems from the last to the first time period, so the solutions of future periods can be

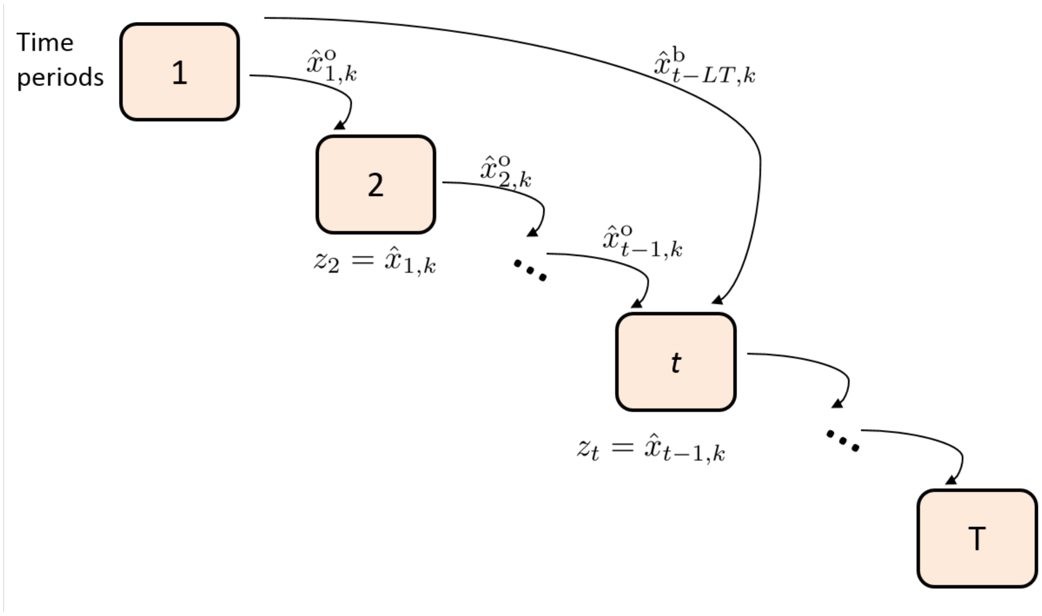

Figure 5.4: Forward Pass for iteration k generates a feasible solution to the full/original problem (over the full planning horizon)

used to generate cuts to provide approximations to predict the cost-to-go functions within the planning horizon. These are cumulative cuts, but specific for each time period; i.e., they are added at each iteration k whenever a new Backward Pass subproblem for year t is solved, and they are kept in the formulation of the following Forward Pass. Note that the fixed variables stored in the Forward Pass, $\hat{x}_t$, are also used in the Backward Pass. The overall procedure of the Backward Pass is represented in Figure 5.5.

The lower bound, LB_k, is calculated in the Backward Pass as in (5.57). It is easy to see that the relaxed solution of the first time period $t = 1$ is a lower bound to the total cost since it only has a subset of the original constraints of the original problem.

$$LB_k = \hat{\Phi}_{1,k} \qquad \forall\ k \tag{5.57}$$

If our problem were convex and solved by standard Nested Benders Decomposition, the objective value and the Lagrange multiplier of the equality constraint (5.53b) would be

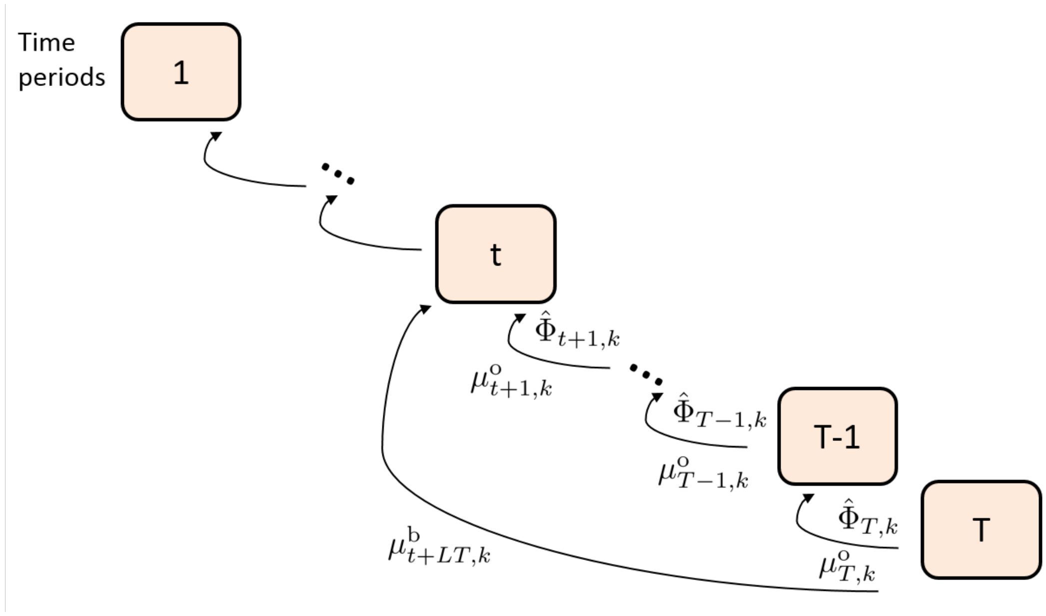

Figure 5.5: The backward pass generates cuts for the cost-to-go function approximations using the solutions from the forward pass

enough to generate the Benders cut (5.58). We assume here that t is fixed.

$$\alpha_t \geq \hat{\Phi}_{t+1,k} + \mu^{\intercal}_{t+1,k}(\hat{x}_{t,k} - x_t) \qquad \forall\ k \tag{5.58}$$

However, the cut cannot be directly generated because the subproblems have integer variables, and thus, are non-convex. In order to provide a valid cut, the subproblems have to be convexified, which can be done by considering the linear or the Lagrangean relaxation of the MILP. The cuts generated by the relaxed problems are the Benders and Lagrangean cuts, respectively (which follows the same notation as Zou et al. (2018b)). A third type of cut proposed by Zou et al. (2018b) as a Strengthened Benders cut is also valid for the Backward Pass subproblems.

The choice of cuts directly impacts the performance of the algorithm as some cuts are tighter and more/less computationally expensive to generate than the others. It is possible, and sometimes recommended, to combine different cuts in the algorithm if one does not strictly dominate the other. Following, we define the different types of cuts that can be used in the Backward Pass, present their potential advantages and disadvantages, and in Section

5.4.3 we demonstrate their validity.

Benders cut The Benders cut is exactly like (5.58), but the coefficients, i.e., the optimal solution of the immediately after period, $\hat{\Phi}^{\mathrm{LP}}_{t+1,k}$, and the multipliers for the complicating equalities, $\mu_{t+1,k}$, are obtained from the solution of the linear relaxation, as represented in equation (5.59). We assume here that t is fixed.

$$\alpha_t \geq \hat{\Phi}^{\mathrm{LP}}_{t+1,k} + \mu^{\intercal}_{t+1,k}(\hat{x}_{t,k} - x_t) \qquad \forall\ k \tag{5.59}$$

This is the weakest of the possible cuts, but it has the advantage of being easily and quickly computed. It is expected to perform well if the formulation is tight and the solution of the linear relaxation is close to the actual solution of the MILP (Sahinidis and Grossmann, 1991; Rahmaniani et al., 2016). For certain multistage capacity planning problems with integer recourse, there is evidence that Benders cuts alone are sufficient for driving the optimality gap to zero as the number of stages increases (see, e.g., Huang and Ahmed (2009), Corollaries 1 and 2). However, it is important to highlight that if this is the cut being used, the algorithm does not have guaranteed finite convergence since there can be a duality gap.

Lagrangean cut The MILP subproblem ($\mathscr{P}_{t,k}$), given by (5.53), can also be convexified by considering its Lagrangean relaxation, which yields the convex hull of the noncomplicating constraint (Frangioni, 2005). In our case this is done by dualizing the linking equalities (5.53b) and penalizing their violation in the objective function by the vector of Lagrange multipliers, $\mu_{t,k}$. Thus, the Lagrangean relaxation of ($\mathscr{P}_{t,k}$) is defined by ($\mathscr{L}_{t,k}$), in (5.60).

$$\begin{aligned} \mathscr{L}_{t,k} : \Phi^{\mathrm{LR}}_t(\mu_{t,k}, \hat{x}_{t-1,k}, \phi_{t,k}) = \quad & \min_{x_t,y_t,z_t} \quad f_t(x_t, y_t) + \phi_{t,k}(\hat{x}_{t,k}) - \mu^{\intercal}_{t,k}(z_t - \hat{x}_{t-1,k}) \\ & \text{s.t.} \qquad (x_t, y_t, z_t) \in \mathcal{X}_t \end{aligned} \tag{5.60}$$

The closer the Lagrange multipliers are to their optimal value, the tighter the approximation is, and the stronger the cuts generated by theses multipliers are. The optimal Lagrange multipliers, $\bar{\mu}_{t,k}$, are obtained by the maximization problem (5.61).

$$\Phi^{\text{LD}}_{t,k}(\hat{x}_{t-1,k}, \phi_{t,k}) = \max_{\mu_{t,k}} \ \Phi^{\text{LR}}_{t,k}(\mu_{t,k}, \hat{x}_{t-1,k}, \phi_{t,k}) \tag{5.61}$$

The Lagrangean cut uses the coefficients obtained by the maximization problem (5.61), as represented in equation (5.62). We assume here that t is fixed.

$$\alpha_t \geq \hat{\Phi}^{\text{LD}}_{t+1,k} + \bar{\mu}^{\intercal}_{t+1,k}(\hat{x}_{t,k} - x_t) \qquad \forall\ k \tag{5.62}$$

The maximization problem in (5.61) can, however, be computationally expensive. Therefore, we adapted the Lagrange multiplier optimization algorithm for each of the subproblems of the Backward Pass based on Thome et al. (2013) using the subgradient method (Fisher, 2004). The steps of the Backward Pass within the Nested Decomposition algorithm if only Lagrangean cuts are used is described below.

For a time period $t = T, ..., 1$ in iteration k:

1. Solve the original MILP subproblem in (5.53) to get the actual objective value, $\Phi^{\text{OP}}_{t,k}$;
2. Solve the linear relaxation of the MILP subproblem in (5.53), and store the dual variables, $\mu_{t,k}$;
3. Use the dual variables from the LP relaxation as an initial guess for the Lagrange multipliers;
4. Solve the Lagrangean subproblem in (5.60) to obtain the optimal value $\Phi^{\text{LR}}_{t,k}$;
5. Check stopping criteria:
    (a) If $\Phi^{\text{OP}}_{t,k} - \Phi^{\text{LR}}_{t,k} \leq \epsilon_2$, where ϵ_2 is a pre-specified tolerance, store the optimal value $\Phi^{\text{LR}}_{t,k}$, and multipliers $\mu_{t,k}$, and go to the next subproblem $t-1$, adding the appropriate future cost cut.
    (b) If no significant progress is achieved after re-solving the Lagrangean relaxation, i.e., if $|\Phi^{\text{LR,old}}_{t,k} - \Phi^{\text{LR}}_{t,k}| \leq \epsilon_3$, where ϵ_3 is a pre-specified tolerance and $\Phi^{\text{LR,old}}_{t,k}$ is the solution of the Lagrangean Relaxation in the

previous step of the subgradient method, this means that no further effort should be made to decrease the duality gap of this subproblem in this iteration k. Store the optimal value $\Phi^{\text{LR}}_{t,k}$, and the multipliers $\mu_{t,k}$, and go to the next subproblem $t-1$, adding the appropriate future cost cut.

6. If the stopping criteria are not met, update the set of multipliers using the subgradient method:
   $\mu_{t,k} = \mu_{t,k} - step_{t,k} \cdot (z_t - \hat{x}_{t-1,k})$

   where $step_{t,k} = \frac{\Phi^{\text{OP}}_{t,k} - \Phi^{\text{LR}}_{t,k}}{(z_t)^2}$, and go back to step 3.

Zou et al. (2018b) proved that if all the linking variables are binary, the Lagrangean cut is tight and the Nested Decomposition algorithm converges in a finite number of steps. However, in our case the state variables are integer and continuous, thus finite convergence is not guaranteed because there may be a duality gap associated with the solution (Ahmed, 2016).

Strengthened Benders cut As mentioned in the previous sections, depending on the structure of the problem, the Benders cuts can be weak and the Lagrangean cuts can take a long time to compute. In order to mitigate potential performance issues, Zou et al. (2018b) proposed the Strengthened Benders cut, which is a compromise between Benders and Lagrangean cuts. Its generation is similar to the Lagrangean cut, but it does not use the subgradient method to improve the multipliers. Instead, it uses the coefficients from the first Lagrangean relaxation solved after the initialization of the multipliers using LP relaxation as shown in (5.63). We assume that t is fixed here.

$$\alpha_t \geq \hat{\Phi}^{\text{LR}}_{t+1,k} + \mu^{\intercal}_{t+1,k}(\hat{x}_{t,k} - x_t) \qquad \forall\ k \tag{5.63}$$

These cuts are at least as tight as the Benders cut (Zou et al., 2018b), but usually less computationally expensive than the Lagrangean cuts.

5.4.3 Validity of the cuts

The Nested Decomposition algorithm is valid as long as the cuts generated in the Backwards Pass are valid according to the following definition (Zou et al., 2018b).

Definition 1. Let $\{(\hat{\Phi}_{t,k}, \mu_{t,k})\}$ be the coefficients obtained from the Backward Pass in iteration k, and $\hat{x}_{t,k}$ be the value for x_t obtained in the Forward Pass of iteration k and fixed for the Backward Pass. Let $\Phi_t(\hat{x}_{t-1,k})$ be the optimal solution for subproblem t for a $\hat{x}_{t-1,k}$ assuming exact representation of the cost-to-go-function. A cut is valid for all periods $t \in \mathcal{T}$ and all iterations $k \in \mathcal{K}$ if

$$\Phi_t(\hat{x}_{t-1,k}) \geq \hat{\Phi}_{t+1,k} + (\mu_{t+1,k})^{\intercal}(\hat{x}_{t,k} - x_t) \tag{5.64}$$

The validity of the Benders, Lagrangean and Strengthened Benders cuts in the context of mixed-integer state variables is proved in Propositions 3, 4, and 5, respectively. Propositions 4 and 5 readily follows from validity part of Theorem 2 in Zou et al. (2018b), but they now allow mixed-integer state variables.

Proposition 3. The Benders cut (5.59), generated by solving the linear relaxation of $(\mathscr{P}_{t,k})$, underestimates $\Phi_t(\hat{x}_{t-1,k})$.

Proof. It trivially follows that the linear relaxation of $(\mathscr{P}_{t,k})$, with optimal value $\Phi^{\mathrm{LP}}_{t,k}$, underestimates the optimal value $\Phi_{t,k}$ of the original problem $(\mathscr{P}_{t,k})$, that is, $\Phi_{t,k} \geq \Phi^{\mathrm{LP}}_{t,k}$. Therefore, since the Benders cut is valid for the LP problem, as proved by Birge (1985), it is also valid for the MILP problem. □

Proposition 4. The Lagrangean cut (5.62), generated by solving the Lagrangean dual of $(\mathscr{P}_{t,k})$, underestimates $\Phi_t(\hat{x}_{t-1,k})$.

Proof. This proof follows from the proof presented in Zou et al. (2018b), however in latter only binary state variables are considered, and now we allow mixed-integer state variables. For the last period $t = T$, the cost-to-go function $\phi_{t,k} = 0$. If we relax the equality $z_t = \hat{x}_{t-1,k}$

in $(\mathscr{P}_{t,k})$ using the optimal multiplier of the Lagrangean problem (5.61), $\bar{\mu}_{k,t}$, we have for any $\hat{x}_{t-1,k} \in \mathbb{Z}_+^{n_1} \times \mathbb{R}_+^{n_2}$, where $n_1 + n_2 = n$,

$$\begin{aligned} \Phi_t(\hat{x}_{t-1,k}) \geq & \min_{x_t, y_t, z_t} \left\{ f_t(x_t, y_t) - \bar{\mu}_{t,k}^\intercal (z_t - \hat{x}_{t-1,k}) : (x_t, y_t, z_t) \in \mathcal{X}_t \right\} \\ = & \quad \Phi_{t,k}^{\mathrm{LD}} \end{aligned} \tag{5.65}$$

Thus, the Lagrangean cut is valid for $t = T$. Next, we prove by induction that the Lagrangean cut is also valid for the remaining timeperiods. Consider $t \leq T - 1$ and assume the Lagrangean cut defined by $\{(\hat{\Phi}_{t+2,k}, \bar{\mu}_{t+2,k}, \hat{x}_{t+1,k})\}$ is valid. Note that:

$$\Phi_t(\hat{x}_{t-1,k}) = \min_{x_t, y_t, z_t, \alpha_t} \left\{ f_t(x_t, y_t) + \alpha_t : (x_t, y_t, z_t) \in \mathcal{X}_t,\ z_t = \hat{x}_{t-1,k},\ \ \alpha_t \geq \Phi_{t+1} \right\} \tag{5.66}$$

Since the cut defined by $\{(\hat{\Phi}_{t+2,k}, \bar{\mu}_{t+2,k}, \hat{x}_{t+1,k})\}$ is valid, i.e. $\Phi_{t+1}(x_t) \geq \hat{\Phi}_{t+2,k} + (\bar{\mu}_{t+2,k})^\intercal (\hat{x}_{t+1,k} - x_{t+1})$ for any $x_{t+1} \in \mathbb{Z}_+^{n_1} \times \mathbb{R}_+^{n_2}$, then the new feasible region $\mathcal{X}_t'$ that incorporates this cut is a relaxation of the original feasible region $\mathcal{X}_t$ of (5.66). Hence, by dualizing the equality constraint we have that

$$\begin{aligned} \Phi_t(\hat{x}_{t-1,k}) \geq & \min_{x_t, y_t, z_t, \alpha_t} \left\{ f_t(x_t, y_t) + \alpha_t : \quad (x_t, y_t, z_t) \in \mathcal{X}_t', z_t = \hat{x}_{t-1,k} \right\} \\ \geq & \min_{x_t, y_t, z_t, \alpha_t} \left\{ f_t(x_t, y_t) + \alpha_t - \quad \bar{\mu}_{t,k}^\intercal (z_t - \hat{x}_{t-1,k}) : (x_t, y_t, z_t) \in \mathcal{X}_t' \right\} \\ = & \quad \Phi_{t,k}^{\mathrm{LD}} \end{aligned} \tag{5.67}$$

Thus, the Lagrangean cut defined by $\{(\hat{\Phi}_{t+1,k}, \bar{\mu}_{t+1,k}, \hat{x}_{t,k})\}$ is valid for $t \in \mathcal{T}$, which completes the proof of Proposition 4. □

Proposition 5. The Strengthened Benders cut (5.63), generated by solving the Lagrangean relaxation of $(\mathscr{P}_{t,k})$ using the multipliers from the linear relaxation of $(\mathscr{P}_{t,k})$, underestimates $\Phi_t(\hat{x}_{t-1,k})$.

Proof. Since the Lagrangean dual is the tightest of the Lagrangean relaxations, the proof for Proposition 4 is also valid for Proposition 5. □

5.4.4 Accelerated Nested Decomposition Algorithm

The proposed Nested Decomposition algorithm (as described in Sections 5.4.2 and 5.4.2) performs very well for our type of problem as will be shown in the results. However, there is potential for improvement, particularly in the early iterations, since the initial planning years have little information of what happens in future years. It is well known that Benders decomposition and its variants can oscillate wildly during initial iterations when the cost-to-go approximations of future stages is poor (Rahmaniani et al., 2016). Here, we propose an acceleration technique aimed at 'warm-starting' the initial cost-to-go approximations with information from an aggregated expansion model, which can potentially speed up the convergence.

The first step is to solve an aggregated version of the full-space MILP and use its solution to pre-generate cuts before entering in the first Forward Pass. The level of aggregation can be decided by the user. The key is to balance the trade-off between a highly aggregated model, which is fast to solve but generates weaker cuts, and a barely aggregated model, which will take almost as long to solve as the original MILP (especially if the solution of the LP relaxation is the main bottleneck) but generates stronger cuts. After some preliminary experiments, we opted to aggregate the model by having only one representative day per year with hourly-level information and relaxing the integrality of the unit commitment variables.

The aggregated model can provide multiple solutions for the linking variables x_t by using the solver's *solution pool* option (IBM, 2015). The decision of how many solutions to use for cut generation is also user-defined (the larger the number of solutions used, the better the representation of the original model, but the longer it takes to compute). These solution values are fixed, $\hat{x}_{t,0}$, and used to generate cuts in a pre-Backward Pass, which is solved before entering the Nested Decomposition iterations. Thus, the algorithm uses the information from this aggregated model that has a view of the full planning horizon (but in an approximated fashion) to gather information about future periods and what are potential good solutions, so that in early iterations, the cost-to-go approximations possess stronger cuts and, consequently, more relevant information about the cost of future investment decisions. Note that any of the cuts presented before can be used in this step of the algorithm.

After finishing this pre-Backward Pass, the algorithm goes to the first Forward Pass keeping the pre-generated cuts. From this point on, the algorithm follows the same steps as in the standard Nested Decomposition. The main steps of the improved Nested Decomposition are shown in Figure 5.6.

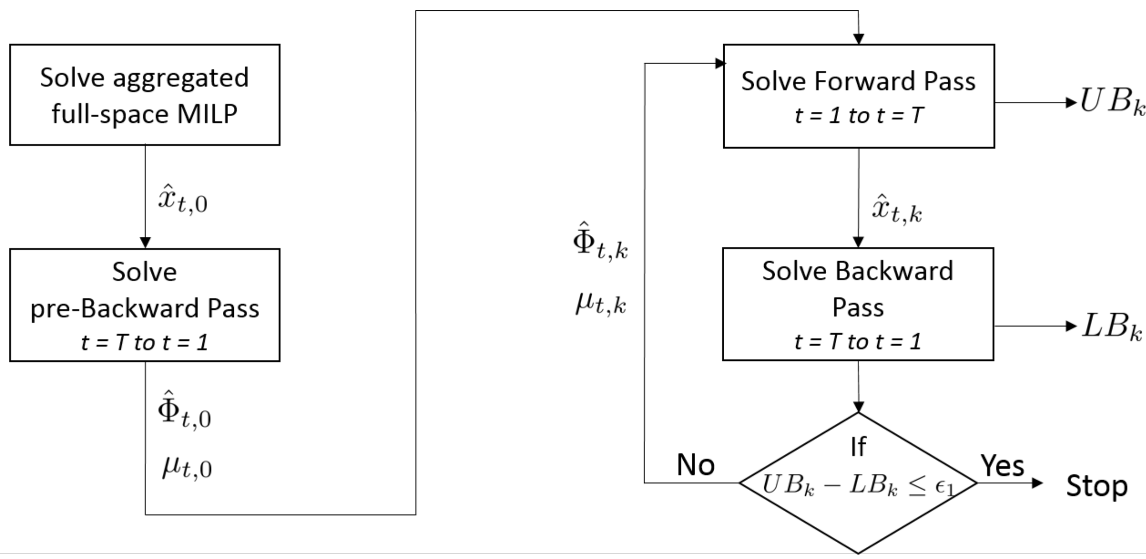

Figure 5.6: Accelerated Nested Decomposition algorithm using pre-generated cuts from an aggregated model to improve initial cost-to-go approximations.

5.5 Case study

In order to test the formulation proposed in Section 5.3 and the algorithm proposed in Section 5.4, we applied them to a case study approximating the Texas Interconnection, a power grid that covers most of the state of Texas. This system is managed by the Electric Reliability Council of Texas (ERCOT), which is an independent system operator responsible for the flow of electric power of about 90% of Texas' electric load. The choice for this interconnection is based on the fact that it is one of the three minor grids in the continental

U.S. power transmission grid, so it is a fairly isolated system that has manageable size. Since our focus is on assessing our multi-scale model and decomposition algorithms, the results presented here should not be interpreted as a detailed analysis of the ERCOT system along the lines of other efforts in the literature (Newell et al., 2014).

Within the ERCOT covered area, we consider four geographical regions: Northeast, West, Coastal and South. We also include a fifth region, Panhandle, which is technically outside the ERCOT limits but due to its renewable generation potential, it supplies electricity to the ERCOT regions. Thus, Panhandle is considered a zone with zero load demand, i.e., it is only a supplier, not a consumer. The regions are shown in Figure 5.1. We assume the geographical midpoints for the Northeast, West, Coastal, South and Panhandle are Dallas, Glasscock County, Houston, San Antonio, and Amarillo, respectively.

For each of the regions, we use load and capacity factor profiles with an hourly resolution. Representative days are constructed using a k-means clustering algorithm. After normalizing 2004-2010 zonal load and renewables profiles into a list of 2555 vectors (7 years of data times 365 days per year – leap days were excluded), d clusters (for $d = 1, ..., D = 12$) are constructed to find a most representative day, chosen as the day closest in Euclidean distance to the centroid of each resulting cluster. Each representative day is then assigned a weight proportional to the number of historical days in the corresponding cluster. Finally, the weighted load profile is normalized to equal 2015 aggregate ERCOT load of 347.5 TWh so that a fair comparison between runs with different numbers of clusters can be made. We assume a constant load growth of 1.4%/year to project load for the 30 years of the planning horizon being studied.

The investment cost, fixed and variable operating costs are derived from the National Renewable Energy Laboratory (NREL), available in the 2016 Annual Technology Baseline (ATB) Spreadsheet (NREL, 2016). We consider a 30-year time horizon, in which the first year is 2015. The fuel price data for coal, natural gas and uranium correspond to the reference scenario of EIA Annual Energy Outlook 2016 (EIA, 2016). A discount rate of 5.7% as chosen in Short et al. (2011) is used to distinguish between the costs incurred in various years of the planning horizon. All of our computation results assume no carbon tax or renewable generation quota is imposed, and no storage expansion allowed.

We assume that the transmission loss is 1%/100 miles, which is the same fraction loss per miles used by Short et al. (2011). The transmission line capacities were generated combining information from Park and Baldick (2013); ERCOT Systems Planning (2014); ABB (2010). The operating data such as ramping limits, minimum operating reserve requirements, are from a range of sources (Mai et al., 2013; Kerl et al., 2015; NWPCC, 2010; WEC, 1984; ERCOT, 2016c). The planning reserve requirement is assumed to be 13.75% (Potomac Economics, 2016). For the nominal capacity of the clusters in each region, we use data from ERCOT (2016c), which has a list of all the power plants in ERCOT for different categories. We compile this data into 7 existing cluster types and combine them with the data from ERCOT (2016) to have this information divided by the regions. Then, for each of the clusters in each region, we assume that the nominal capacity is the mean size of all generators within that cluster.

We first solve the proposed model for 1 to 12 representative days, with an optimality gap tolerance of 1%, to assess the impact of the number of representative days in the planning strategy. Then, we solve the proposed model as a full-space MILP, as well as using the Nested Decomposition with the cuts presented in Section 5.4.2 for the 4 representative days variant of the model, and applied the acceleration technique from Section 5.4.4 to the cut with the best performance. The two best versions of the algorithm are also tested for the 12 representative days.

Our computational tests were performed on a standard desktop computer with an Intel(R) Core(TM) i7-2600 CPU @ 3.40 GHz processor, with 8GB of RAM, running on Windows 7. We implement the monolithic formulation and all the versions of the Nested Decomposition algorithm in GAMS 24.7.1, and solve the LPs and MILPs using CPLEX version 12.6.3.

5.5.1 Impact of the number of representative days in the planning strategy

To evaluate the impact of including more representative days on the generation investment decisions, we ran the proposed model for 1 to 12 representative days per year. The fractional difference relative to the 12-day representation by generation source is shown in Figure 5.7. This plot does not include coal and nuclear because their final year capacities are exactly the same irrespective of the number of representative days considered. The wind capacity suffers minor fluctuation and slowly converges to a capacity that is 3% higher than the 1-day model projected. On the other hand, a major difference occurs in projected PV and NG generation capacities. For 1 to 4 representative days, there is a clear trend of decreasing PV and increasing NG contributions as we increase the number of representative days. With the exception of the outlier in the "5 representative days" results, this trend can be seen for all the variants until the results reach a plateau with small fluctuations. This trend is in agreement with the results presented in Nahmmacher et al. (2016b). It is important to highlight that the jump in the installed capacity of PV and natural gas with 5 representative days is due in part to the approach for selecting representative days. Namely, historical data are clustered according to the *joint* distribution of load and renewables capacity factors. This does not guarantee that the profile used to represent individual technologies will better match that technology's historical duration curves as the number of representative days is increased.

After this preliminary experiment, we opted to test the algorithm for 4 and 12 representative days per year. The 4 representative days variant was chosen because its results are similar to those of the 12-day variant, while requiring less computational time. We also opted to solve the 12 representative days variant because it exploits better the full potential of the Nested Decomposition algorithm, which may be needed for more complex models that require higher time resolution (e.g, models that consider storage) Nahmmacher et al. (2016b).

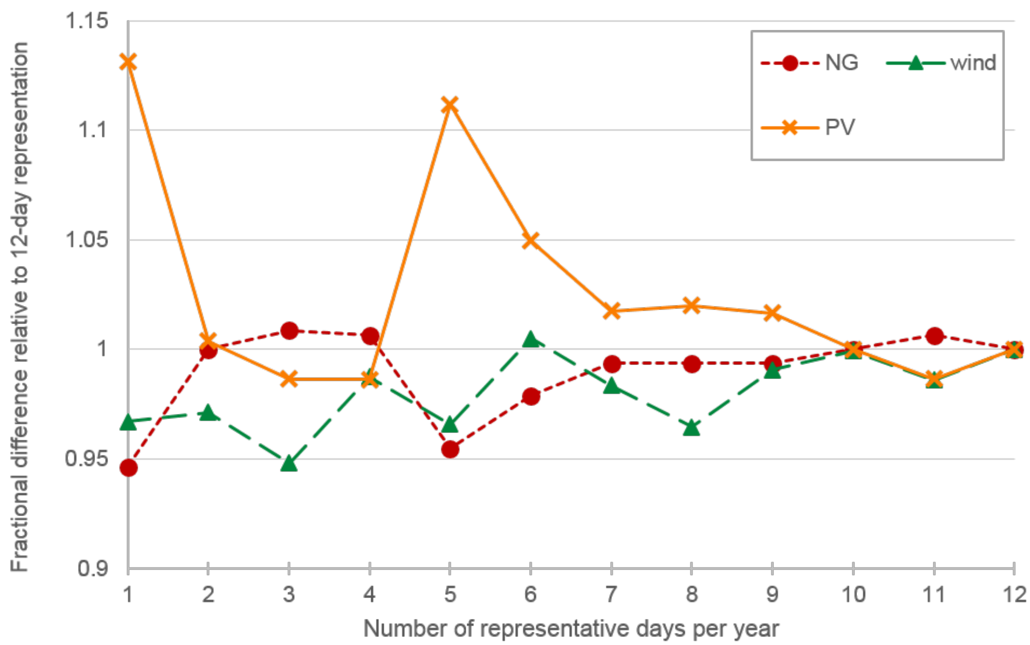

Figure 5.7: Capacity projections of natural gas, solar PV and wind at the end of the time horizon, varying with the number of representative days selected. The capacities are represented as a fraction of capacity projected by the 12-day model.

5.5.2 4 representative days variant

The full-space MILP model using 4 representative days has 1,201,761 constraints, 413,644 discrete variables, and 594,147 continuous variables. After invoking CPLEX's presolver, the model was reduced to 800,755 constraints, 25,231 binary variables, 388,493 integer variables, and 225,367 continuous variables. We solved the problem using several methods: (1) we solved the monolithic formulation directly in CPLEX; (2) we applied the Nested Decomposition algorithm with each of the cut options; (3) we employed the acceleration technique to the Nested Decomposition variant with the best performance in (2). For all the versions of the algorithm, we set a maximum of 20 iterations and an optimality gap tolerance of 0.01% for any MILP subproblem, and for the full-space MILP we set a maximum time of 4 hours. The comparison between the performance of the full-space MILP and the various algorithm versions is shown in Figure 5.8.

All of the versions of the algorithm found solutions within 10% gap in less than 10

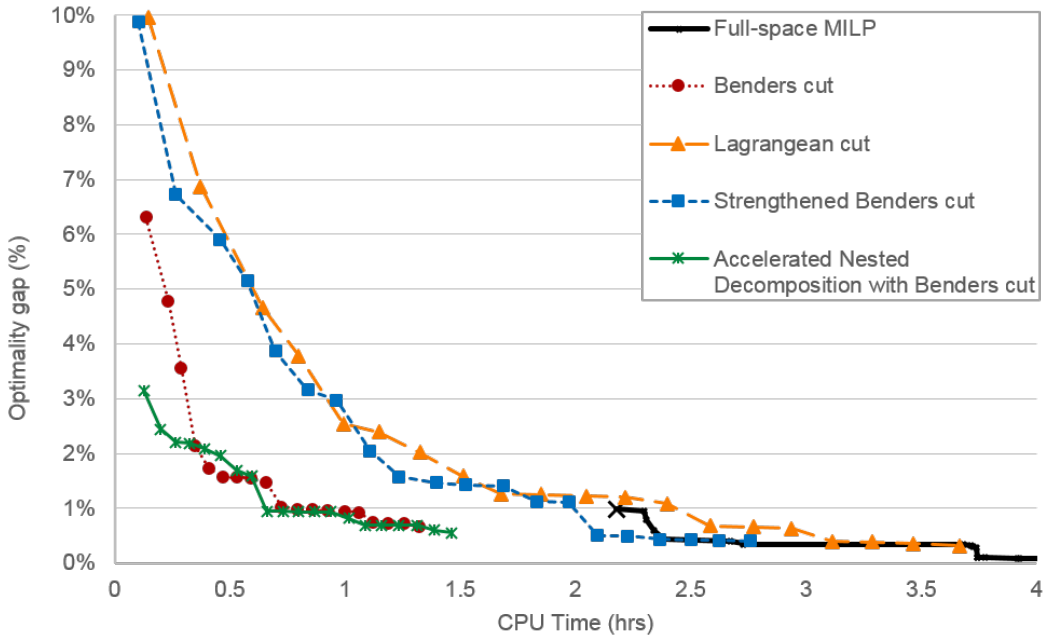

Figure 5.8: Algorithm performance in the 4-representative day model. The results show that the Benders cut and its accelerated version are the fastest in finding a solution within 1% optimality gap.

minutes, while the first integer solution by the monolithic MILP took 2.04 hours. The Benders cuts was the most efficient among the possible cuts, finding a solution within 1% gap in 47.5 minutes. This was already expected due to the tightness of the linear relaxation, and to how quickly the Benders cuts are generated. Further improvements can be gained by employing the accelerated version (see Section 5.4.4) in which warm-start cuts are generated before invoking the Nested Benders approach. The accelerated Nested Decomposition with Benders cuts greatly reduces the optimality gap in the initial iterations and is the fastest (39.6 minutes) at finding a solution within 1% optimality gap.

5.5.3 12 representative days variant

The full-space MILP model for the 12 representative days variant has 3,626,721 constraints, 1,243,084 discrete variables, and 1,774,947 continuous variables. After invoking CPLEX's presolver, the model was reduced to 2,428,687 constraints, 77,071 binary vari-

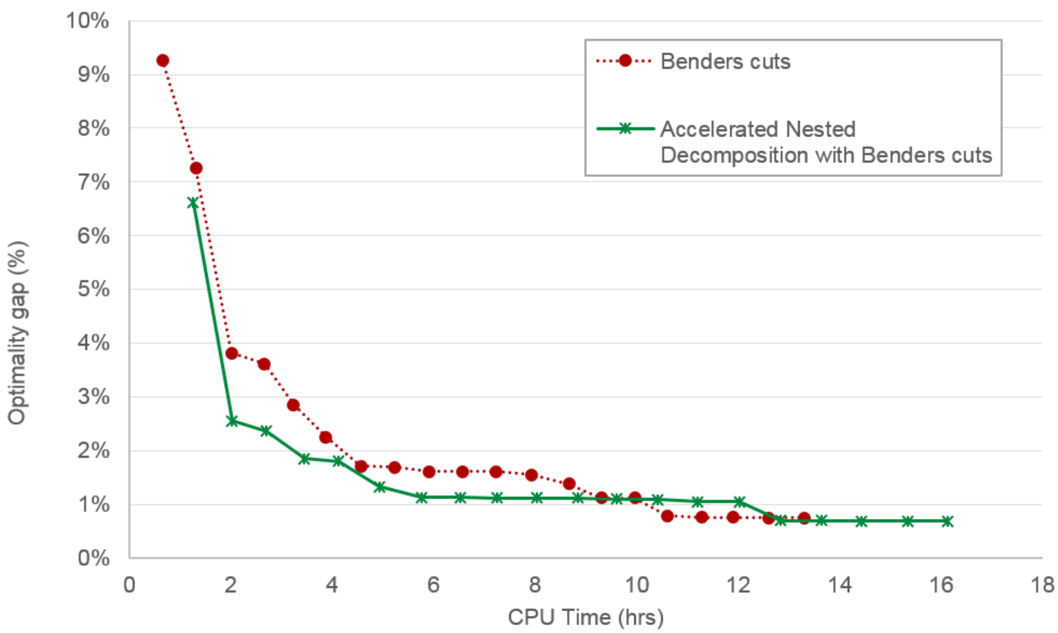

Figure 5.9: Algorithm performance in the 12-representative day model. The results show that the Accelerated Nested Decomposition algorithm provides smaller optimality gaps in the initial iterations.

ables, 1,183,373 integer variables, and 673,825 continuous variables. Despite these reductions, CPLEX terminated with a 100% optimality gap after solving the monolithic formulation for 24 hours, our chosen time limit. We solved the problem using the Nested Decomposition algorithm with the two versions that performed the best using 4 representative days (Benders cuts, and warm-start cuts + Benders cuts). For those, we set a maximum of 20 iterations and an optimality gap tolerance of 0.01% for any MILP subproblem. The comparison between the performance of the two algorithms is shown in Figure 5.9.

Both versions of the algorithm obtained solutions within 2% optimality gap in less than 5 hours. We notice again that by "warm-starting" the algorithm we were able to reduce the optimality gap in the initial iterations. The accelerated version found a solution with a 2% gap in its 4^{th} iteration and 3.5 hours, while the normal Benders cut version took 4.6 hours and 7 iterations to reach the same optimality gap.

5.5.4 Cost breakdown comparison

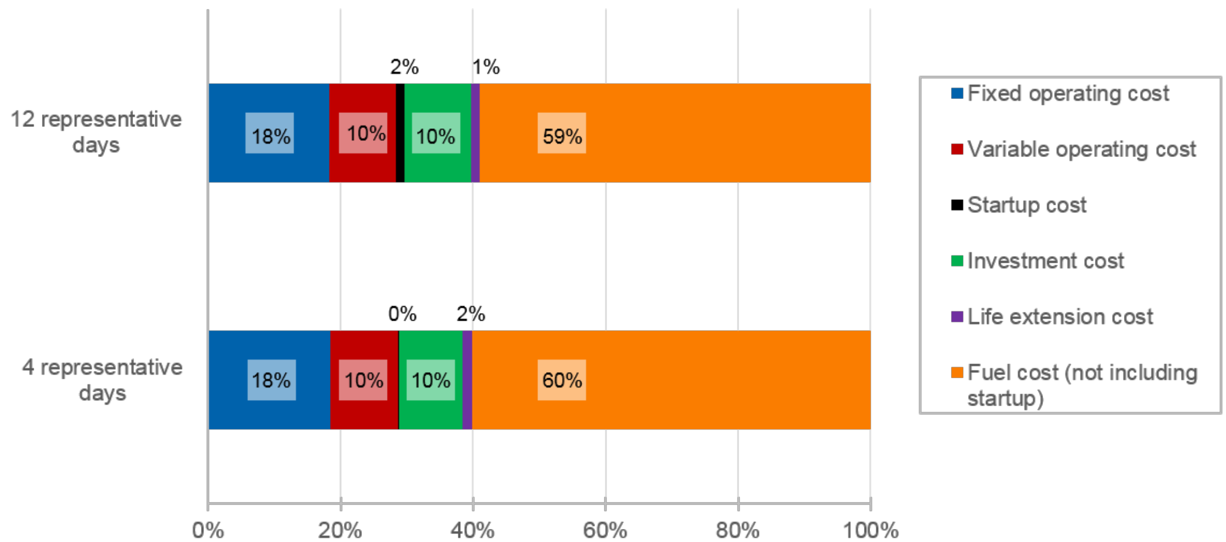

Figure 5.10: Breakdown of system costs using 4 and 12 representative days.

We also compare the impact of the having more representative days in the cost breakdown, which is shown in Figure 5.10. The results indicate that for both cases the cost is driven by the fuel consumption (coal, natural-gas and uranium), which accounts for about 60% of the total cost. The major difference between these representations is in the startup contribution, which goes from 0.3% to 1.5% of the total cost when considering 4 and 12 representative days. These results show that by having a better representation of the dynamics of the systems, i.e., higher number of representative days with hourly load and capacity factor, the startup cost becomes more relevant.

5.6 Conclusions

In this chapter we propose a multiperiod MILP model to solve power systems planning problem considering increasing share of renewables. We adopt clustering and time scale strategies in order to reduce the size of the model without greatly impacting the quality of the results. The major novelties and contributions of this chapter are in the problem formulation, especially in the description of retirement and handling of multi-scale aspects, and in its solution strategy. We develop a decomposition algorithm based on Nested Benders Decomposition for this multi-period deterministic problem with integer and continuous state variables that does not have guaranteed finite convergence as there is a potential duality gap, although in the implementation a maximum number of iterations is specified. We also present several options of valid cuts for the Backward Pass, and make use of an acceleration technique to speed up the conversion of the algorithm.

The formulation and solution framework are tested for a case study in the ERCOT region. The results show that the algorithm can provide substantial speed-up and allow the solution of larger instances. This improvement in solution time is important because it allows one to perform several sensitivity analysis and better understand the drivers for a variety of scenarios. Additionally, we conclude that for the proposed model and case study, it is sufficient to have 4 representative days per year in order to adequately represent the variability in load and capacity factor for the different regions. In other instances, such as when studying the potential role of energy storage technologies, it may be necessary to consider more than 4 representative days to model grid operations. For those cases, the developed algorithms are shown to be critical to finding an optimal solution in a reasonable amount of time.

There are several future research directions worth investigating. Regarding the model formulation, it would be interesting to see the impacts of including transmission expansion, having a more detailed representation of transmission, and incorporating long-term uncertainty into the model. Regarding the algorithm, there is potential for parallelization, especially for the Lagrangean cut version of the algorithm, which could improve its performance and make it more competitive with other versions.

Nomenclature

Indices and Sets

$r \in \mathcal{R}$ set of regions within the area considered

$i \in \mathcal{I}$	set of generator clusters
$i \in \mathcal{I}_r$	set of generator clusters in region r
$i \in \mathcal{I}_r^{\text{old}}$	set of existing generator clusters in region r at the beginning of the time horizon, $\mathcal{I}_r^{\text{old}} \subseteq \mathcal{I}_r$
$i \in \mathcal{I}_r^{\text{new}}$	set of potential generator clusters in region r, $\mathcal{I}_r^{\text{new}} \subseteq \mathcal{I}_r$
$i \in \mathcal{I}_r^{\text{TH}}$	set of thermal generator clusters in region r, $\mathcal{I}_r^{\text{TH}} \subseteq \mathcal{I}_r$
$i \in \mathcal{I}_r^{\text{RN}}$	set of renewable generator clusters in region r, $\mathcal{I}_r^{\text{RN}} \subseteq \mathcal{I}_r$
$i \in \mathcal{I}_r^{\text{Told}}$	set of existing thermal generator clusters in region r, $\mathcal{I}_r^{\text{Told}} \subseteq \mathcal{I}_r^{\text{TH}}$
$i \in \mathcal{I}_r^{\text{Tnew}}$	set of potential thermal generator clusters in region r, $\mathcal{I}_r^{\text{Tnew}} \subseteq \mathcal{I}_r^{\text{TH}}$
$i \in \mathcal{I}_r^{\text{Rold}}$	set of existing renewable generator clusters in region r, $\mathcal{I}_r^{\text{Rold}} \subseteq \mathcal{I}_r^{\text{RN}}$
$i \in \mathcal{I}_r^{\text{Rnew}}$	set of potential renewable generator clusters in region r, $\mathcal{I}_r^{\text{Rnew}} \subseteq \mathcal{I}_r^{\text{RN}}$
$j \in \mathcal{J}$	set of storage unit clusters
$t \in \mathcal{T}$	set of time periods (years) within the planning horizon
$d \in \mathcal{D}$	set of representative days in each year t
$s \in \mathcal{S}$	set of sub-periods of time per representative day d in year t
$k \in \mathcal{K}$	set of iterations in the Nested Decomposition algorithm

Deterministic Parameters

$L_{r,t,d,s}$	load demand in region r in sub-period s of representative day d of year t (MW)
$L_t^{\max}$	peak load in year t (MW)
W_d	weight of the representative day d
Hs	duration of sub-period s (hours)
$Qg_{i,r}^{\text{np}}$	nameplate (nominal) capacity of a generator in cluster i in region r (MW)
$Ng_{i,r}^{\text{old}}$	number of existing generators in each cluster, $i \in \mathcal{I}_r^{\text{old}}$, per region r at the beginning of the time horizon
$Ng_i^{\max}$	maximum number of generators in the potential clusters $i \in \mathcal{I}_r^{\text{new}}$
$Q_{i,t}^{\text{inst,UB}}$	upper bound on yearly capacity installations based on generation technology $(^{\text{MW}}/_{\text{year}})$
$R_t^{\min}$	system's minimum reserve margin for year t (fraction of the peak load)
ED_t	energy demand during year t (MWh)
LT_i	expected lifetime of generation cluster i (years)
T_t^{remain}	remaining time until the end of the time horizon at year t (years)
$Ng_{i,r,t}^{\text{r}}$	number of generators in cluster i of region r that achieved their expected lifetime
Q_i^{v}	capacity value of generation cluster i (fraction of the nameplate capacity)
$Cf_{i,r,t,d,s}$	capacity factor of generation cluster $i \in \mathcal{I}_r^{\text{RN}}$ in region r at sub-period s, of representative day d of year t (fraction of the nameplate capacity)
$Pg_i^{\min}$	minimum operating output of a generator in cluster $i \in \mathcal{I}_r^{\text{TH}}$ (fraction of the nameplate capacity)
$Ru_i^{\max}$	maximum ramp-up rate for cluster $i \in \mathcal{I}_r^{\text{TH}}$ (fraction of nameplate capacity)
$Rd_i^{\max}$	maximum ramp-down rate for cluster $i \in \mathcal{I}_r^{\text{TH}}$ (fraction of nameplate capacity)
F_i^{start}	fuel usage at startup (MMbtu/MW)

$Frac_i^{\mathrm{spin}}$	maximum fraction of nameplate capacity of each generator that can contribute to spinning reserves (fraction of nameplate capacity)
$Frac_i^{\mathrm{Qstart}}$	maximum fraction of nameplate capacity of each generator that can contribute to quick-start reserves (fraction of nameplate capacity)
Op^{min}	minimum total operating reserve (fraction of the load demand)
$Spin^{\mathrm{min}}$	minimum spinning operating reserve (fraction of the load demand)
$Qstart^{\mathrm{min}}$	minimum quick-start operating reserve (fraction of the load demand)
α^{RN}	fraction of the renewable generation output covered by quick-start reserve (fraction of total renewable power output)
$T_{r,r'}^{\mathrm{loss}}$	transmission loss factor between region r and region $r' \neq r$ (%/miles)
$D_{r,r'}$	distance between region r and region $r' \neq r$ (miles)
$Ns_{j,r}$	number of existing storage units in each cluster j per region r at the beginning of the time horizon
$Charge_j^{\mathrm{min}}$	minimum operating charge for storage unit in cluster j (MW)
$Charge_j^{\mathrm{max}}$	maximum operating charge for storage unit in cluster j (MW)
$Discharge_j^{\mathrm{min}}$	minimum operating discharge for storage unit in cluster j (MW)
$Discharge_j^{\mathrm{max}}$	maximum operating discharge for storage unit in cluster j (MW)
$Storage_j^{\mathrm{min}}$	minimum storage capacity for storage unit in cluster j (MWh)
$Storage_j^{\mathrm{max}}$	maximum storage capacity (i.e. nameplate capacity) for storage unit in cluster j (MWh)
η_j^{charge}	charging efficiency of storage unit in cluster j (fraction)
$\eta_j^{\mathrm{discharge}}$	discharging efficiency of storage unit in cluster j (fraction)
LT_j^{s}	lifetime of storage unit in cluster j (years)
Ir	nominal interest rate
If_t	discount factor for year t
$OCC_{i,t}$	overnight capital cost of generator cluster i in year t ($/MW)
$ACC_{i,t}$	annualized capital cost of generator cluster i in year t ($/MW)
$DIC_{i,t}$	discounted investment cost of generator cluster i in year t ($/MW) [2]
$SIC_{j,t}$	investment cost of storage cluster j in year t ($/MW)
CC_i^{m}	capital cost multiplier of generator cluster i (unitless)
LE_i	life extension cost for generator cluster i (fraction of the investment cost of corresponding new generator)
$FOC_{i,t}$	fixed operating cost of generator cluster i ($/MW)
$P_{i,t}^{\mathrm{fuel}}$	price of fuel for generator cluster i in year t ($/MMBtu)
HR_i	heat rate of generator cluster i (MMBtu/MWh)
$Tx_t^{\mathrm{CO_2}}$	carbon tax in year t ($/kg CO_2)
$EF_i^{\mathrm{CO_2}}$	full lifecycle CO_2 emission factor for generator cluster i (kgCO_2/MMBtu)
$VOC_{i,t}$	variable O&M cost of generator cluster i ($/MWh)

[2] $DIC_{i,t}$ is used in the calculation for the life extension investment cost, which is in terms of a fraction LE_i of the capital cost. Therefore the investment cost for the existing cluster is approximated as being the same as for the potential clusters that have the same or similar generation technology.

$RN_t^{\min}$	minimum renewable energy production requirement during year t (fraction of annual energy demand)
PEN_t^{rn}	penalty for not meeting renewable energy quota target during year t ($^{\$}/_{\mathrm{MWh}}$)
PEN_t^{c}	penalty for curtailment during year t ($^{\$}/_{\mathrm{MWh}}$)
C_i^{start}	fixed startup cost for generator cluster i ($^{\$}/_{\mathrm{MW}}$)
$\hat{ngo}_{i,r,t,k}^{\mathrm{rn}}$	solution from the Forward Pass, iteration k, of the number of operational renewable generators in cluster i of region r at year t, $ngo_{i,r,t}^{\mathrm{rn}}$, which is a fixed parameter for the year $t+1$ (unitless)
$\hat{ngo}_{i,r,t,k}^{\mathrm{th}}$	solution from the Forward Pass, iteration k, of the number of operational thermal generators in cluster i of region r at year t, $ngo_{i,r,t}^{\mathrm{th}}$, which is a fixed parameter for the year $t+1$ (unitless)
$\hat{ngb}_{i,r,t,k}^{\mathrm{rn}}$	solution from the Forward Pass, iteration k, of the number of renewable generators built in cluster i of region r at year t, $ngb_{i,r,t}^{\mathrm{rn}}$, which is a fixed parameter for the year $t+LT_i$ (unitless)
$\hat{ngb}_{i,r,t,k}^{\mathrm{th}}$	solution from the Forward Pass, iteration k, of the number of thermal generators built in cluster i of region r at year t, $ngb_{i,r,t}^{\mathrm{th}}$, which is a fixed parameter for the year $t+LT_i$ (unitless)
$\mu_{i,r,t,k}^{\mathrm{o,rn}}$	multiplier of the linking equality related to the number of operational renewable generators in cluster i of region r at year $t-1$ in iteration k (unitless)
$\mu_{i,r,t,k}^{\mathrm{o,th}}$	multiplier of the linking equality related to the number of operational thermal generators in cluster i of region r at year $t-1$ in iteration k (unitless)
$\mu_{i,r,t,k}^{\mathrm{b,rn}}$	multiplier of the linking equality related to the number of renewable generators built in cluster i of region r at year $t-LT_i$ in iteration k (unitless)
$\mu_{i,r,t,k}^{\mathrm{b,th}}$	multiplier of the linking equality related to the number of thermal generators built in cluster i of region r at year $t-LT_i$ in iteration k (unitless)
$\hat{x}_{t,k}$	fixed solution of x_t in iteration k (concise notation)
$\hat{x}_{t,k}^{\mathrm{o}}$	fixed solution of x_t in iteration k corresponding to $\hat{ngo}_{i,r,t,k}^{\mathrm{rn}}$ and $\hat{ngo}_{i,r,t,k}^{\mathrm{th}}$ (concise notation)
$\hat{x}_{t,k}^{\mathrm{b}}$	fixed solution of x_t in iteration k corresponding to $\hat{ngb}_{i,r,t,k}^{\mathrm{rn}}$ and $\hat{ngb}_{i,r,t,k}^{\mathrm{th}}$ (concise notation)
$\hat{\Phi}_{t,k}$	fixed optimal value for subproblem corresponding to period t in iteration k
$\mu_{t,k}$	Lagrange multiplier (concise notation)
$\mu_{t,k}^{\mathrm{o}}$	Lagrange multiplier corresponding to $\mu_{i,r,t,k}^{\mathrm{o,rn}}$ and $\mu_{i,r,t,k}^{\mathrm{o,th}}$ (concise notation)
$\mu_{t,k}^{\mathrm{b}}$	Lagrange multiplier corresponding to $\mu_{i,r,t,k}^{\mathrm{b,rn}}$ and $\mu_{i,r,t,k}^{\mathrm{b,th}}$ (concise notation)
$\bar{\mu}_{t,k}$	optimal Lagrange multiplier (concise notation)
$step_{t,k}$	stepsize used in the subgradient method (unitless)
ϵ_1, ϵ_2, ϵ_3	tolerances for the decomposition algorithm

Continuous variables

Φ	net present cost[3] throughout the time horizon, including amortized investment cost, operational and environmental cost (\$)
Φ_t^{opex}	amortized operating costs in year t (\$)
Φ_t^{capex}	amortized investment costs in year t (\$)
Φ_t^{PEN}	amortized penalty costs in year t (\$)
$p_{i,r,t,d,s}$	power output of generation cluster i in region r during sub-period s of representative day d of year t (MW)
def_t^{rn}	deficit from renewable energy quota target during year t (MWh)
$cu_{r,t,ss,s}$	curtailment slack generation in region r during sub-period s of representative day d of year t (MW)
$p^{\text{flow}}_{r,r',t,d,s}$	power transfer from region r to region $r' \neq r$ during sub-period s of representative day d of year t (MW)
$q^{\text{spin}}_{i,r,t,d,s}$	spinning reserve capacity of generation cluster i in region r during sub-period s of representative day d of year t (MW)
$q^{\text{Qstart}}_{i,r,t,d,s}$	quick-start capacity reserve of generation cluster i in region r during sub-period s of representative day d of year t (MW)
$ngo^{\text{rn}}_{i,r,t}$	number of generators that are operational in cluster $i \in \mathcal{I}_r^{\text{RN}}$ of region r in year t (continuous relaxation)
$ngb^{\text{rn}}_{i,r,t}$	number of generators that are built in cluster $i \in \mathcal{I}_r^{\text{RN}}$ of region r in year t (continuous relaxation)
$ngr^{\text{rn}}_{i,r,t}$	number of generators that retire in cluster $i \in \mathcal{I}_r^{\text{RN}}$ of region r in year t (continuous relaxation)
$nge^{\text{rn}}_{i,r,t}$	number of generators that had their life extended in cluster $i \in \mathcal{I}_r^{\text{RN}}$ of region r in year t (continuous relaxation)
$p^{\text{charge}}_{j,r,t,d,s}$	power being charged to storage cluster j is region r, during sub-period s of representative day d of year t (MW)
$p^{\text{discharge}}_{j,r,t,d,s}$	power being discharged to storage cluster j is region r, during sub-period s of representative day d of year t (MW)
$p^{\text{level}}_{j,r,t,d,s}$	state of charge of storage cluster j is region r, during sub-period s of representative day d of year t (MWh)
$p^{\text{level,0}}_{j,r,t,d}$	state of charge of storage cluster j is region r at hour zero of representative day d of year t (MWh)
$nso_{j,r,t}$	number of storage units that are operational in cluster j of region r in year t (continuous relaxation)
$nsb_{j,r,t}$	number of storage units that are built in cluster j of region r in year t (continuous relaxation)
$nsr_{j,r,t}$	number of storage units that retire in cluster j of region r in year t (continuous relaxation)
Φ_t	objective function value for subproblem t assuming exact representation of the cost-to-go function (\$)
$\Phi_{t,k}$	objective function value for subproblem t in iteration k (\$)

[3] All the costs are in 2015 USD.

$\phi_{t,k}$	cost-to-go function (\$)
α_t	expected future year cost, when calculating the cost for year t (\$)
$\Phi^{\text{LP}}_{t,k}$	net present cost of the linear relaxation of the subproblem for year t in iteration k (\$)
$\Phi^{\text{LR}}_{t,k}$	net present cost of the Lagrangean relaxation of the subproblem for year t in iteration k (\$)
$\Phi^{\text{LD}}_{t,k}$	net present cost of the Lagrangean dual of the subproblem for year t in iteration k (\$)
$\Phi^{\text{OP}}_{t,k}$	net present cost of the original MILP subproblem for year t in iteration k (\$)
$ngo^{\text{rn,prev}}_{i,r,t}$	number of generators that are operational in cluster $i \in \mathcal{I}^{\text{RN}}_r$ of region r in year $t-1$ (continuous relaxation)
$ngb^{\text{rn,LT}}_{i,r,t}$	number of generators that are built in cluster $i \in \mathcal{I}^{\text{RN}}_r$ of region r in year $t - LT_i$ (continuous relaxation)
$ngo^{\text{th,prev}}_{i,r,t}$	number of generators that are operational in cluster $i \in \mathcal{I}^{\text{TH}}_r$ of region r in year $t-1$ (continuous relaxation)
$ngb^{\text{th,prev}}_{i,r,t}$	number of generators that are built in cluster $i \in \mathcal{I}^{\text{TH}}_r$ of region r in year $t - LT_i$ (continuous relaxation)
x_t	state (linking) variables in the concise notation
z_t	duplicated state variables in the concise notation
y_t	local variables in the concise notation

Discrete variables

$ngo^{\text{th}}_{i,r,t}$	number of generators that are operational in cluster $i \in \mathcal{I}^{\text{TH}}_r$ of region r in year t (integer variable)
$ngb^{\text{th}}_{i,r,t}$	number of generators that are built in cluster $i \in \mathcal{I}^{\text{TH}}_r$ of region r in year t (integer variable)
$ngr^{\text{th}}_{i,r,t}$	number of generators that retire in cluster $i \in \mathcal{I}^{\text{TH}}_r$ of region r in year t (integer variable)
$nge^{\text{th}}_{i,r,t}$	number of generators that had their life extended in cluster $i \in \mathcal{I}^{\text{TH}}_r$ of region r in year t (integer variable)
$u_{i,r,t,d,s}$	number of thermal generators ON in cluster $i \in \mathcal{I}_r$ of region r during sub-period s of representative day d of year t (integer variable)
$su_{i,r,t,d,s}$	number of generators starting up in cluster i during sub-period s of representative day d in year t (integer variable)
$sd_{i,r,t,d,s}$	number of generators shutting down in cluster i during sub-period s of representative day d in year t (integer variable)

Acronyms

NG	natural gas
ST	steam turbine
CT	gas-fired combustion turbine

CC	combined cycle
CCS	carbon capture and storage
PV	solar photovoltaic
CSP	concentrated solar panel

Chapter 6

Electric Power Infrastructure Planning under Uncertainty: Stochastic Dual Dynamic Integer Programming (SDDiP) and parallelization scheme

In this chapter, we address the long-term planning of electric power infrastructure under multiscale uncertainty. We propose a Multistage Stochastic Mixed-Integer Programming (MSIP) formulation that is an extension of the deterministic model proposed in chapter 5 and can be of the order of quadrillion variables and constraints. To be able to solve such a large-scale model, we decompose the problem using Stochastic Dual Dynamic Integer Programming (SDDiP) (Zou et al., 2018b) and take advantage of parallel processing to solve it more efficiently.

The proposed GEP model follows four out of the five trends listed in the GEP survey by Babatunde et al. (2018): (i) it handles uncertainty, (ii) it considers renewable energy penetrations and includes short-term operating decisions, (iii) it includes the option of adding energy storage, and (iv) it addresses sustainable issues by having the option of imposing minimum renewable generation quota, maximum CO_2 emissions quota, and/or carbon tax.

The only trend from Babatunde et al. (2018) that this chapter does not address is the deregulation of the power sector.

The major contributions of this chapter are the following: (i) application of SDDiP in the context of GEP optimization with integrated operating decisions; (ii) SDDiP with mixed-integer recourse; (iii) parallelization scheme to solve the SDDiP more efficiently; (iv) application of the model and algorithmic framework to a case-study for Electric Reliability Council of Texas (ERCOT) region considering operational and strategic uncertainty.

The remainder of the chapter is organized as follows: Section 6.1 presents the problem statement, discusses modeling assumptions, proposes a concise representation of the multi-stage mixed-integer linear programming model for GEP optimization under uncertainty (the detailed model is shown in Appendix C.1), and discusses how the scenario tree is generated. Section 6.2 describes the SDDiP algorithm, its framework, major assumptions, and how we parallelize the algorithm. In Section 6.3 we first present the results for a case-study for the ERCOT region, showing the order of magnitude of problems the SDDiP algorithm can solve on a personal computer and the solution time. Then, we compare the first-stage *here-and-now* decisions for the reference case with natural gas price uncertainty, the *no nuclear* case with natural gas price uncertainty, carbon tax uncertainty, and high carbon tax uncertainty, and show the value of stochastic programming for the *no nuclear* case with high carbon tax uncertainty. Finally, in Section 6.4 we draw some conclusions.

6.1 Formulation

The proposed GEP problem involves choosing the optimal investment strategy and operating schedule for the power system in order to meet the projected load demand over the time-horizon for each location, while minimizing the expected net present cost over the scenario tree. This is an extension of the MILP problem proposed by Lara et al. (2018a) to multistage stochastic mixed-integer programming, in order to address uncertainty.

A set of existing and potential generators is given, for which the energy source (nuclear, coal, natural gas, wind or solar)[1] and the generation technology are known.

- For the existing generators we consider: (a) **coal:** steam turbine (coal-st-old); (b) **natural gas:** boiler plants with steam turbine (ng-st-old), combustion turbine (ng-ct-old), and combined-cycle (ng-cc-old); (c) **nuclear:** steam turbine (nuc-st-old); (d) **solar:** photovoltaic (pv-old); (e) **wind:** wind turbine (wind-old);
- For the potential generators we consider: (a) **coal:** without (coal-new) and with carbon capture (coal-ccs-new); (b) **natural gas:** combustion turbine (ng-ct-new), combined-cycle without (ng-cc-new) and with carbon capture (ng-cc-ccs-new); (c) **nuclear:** steam turbine (nuc-st-new); (d) **solar:** photovoltaic (pv-new) and concentrated solar panel (csp-new); (e) **wind:** wind turbine (wind-new);

Also known are: their nameplate (maximum) capacity; lifetime; fixed and variable operating costs; start-up cost (fixed and variable); cost for extending their lifetimes; CO_2 emission factor and carbon tax, if applicable; fuel price; and operating characteristics such as ramp-up/ramp-down rates, operating limits, and contribution to spinning and quick start fraction for thermal generators.

For the case of existing generators, their age at the beginning of the time-horizon and location are also known. For the case of potential generators, the capital cost (which is a linear function of its nameplate capacity), and the maximum yearly installation of each generation technology are also given. Also given is a set of potential storage units, with specified technology (we consider as options lithium-ion, lead-acid, and flow batteries), capital cost, power rating, rated energy capacity, charge and discharge efficiency, and storage lifetime. Additionally, the multiple profiles of projected load demand and renewable availability (capacity factor) are given for each location, as well as the distance between locations, the transmission loss per mile, and the transmission line capacity between locations.

The problem is then to find the optimal "here-and-now" decisions for the first-stage and "wait-and-see" decisions for the remaining stages and respective scenarios regarding: a) location, year, type, and number of generators and storage units to install; b) when to

[1] In this chapter we do not consider hydroelectric power as it is available in very limited amounts in the ERCOT region.

retire generators and storage units; c) whether or not to extend the life of the generators that reach their expected lifetime; d) an approximate operating schedule for each installed generator; and e) the approximate power flow between each location in order to meet the projected demand. The goal is to minimize the expected net present cost over the scenario tree (including operating, investment, and environmental costs). The large-scale strategic level uncertain parameters are user-defined and can be, for example, the yearly fuel price, and any potential carbon tax. The small-scale operating level uncertain parameters (renewable generation availability and load demand) are captured through multiple representative days between investment decisions.

6.1.1 Modeling assumptions

In order to improve tractability and allow the solution of the MSIP model for large areas over a few decades with multiple scenarios, we adopted judicious modeling aggregations and approximations to address the multi-scale aspects, both in its spatial and temporal dimensions (Lara et al., 2018a). In order to significantly reduce computation time, generator clustering (Palmintier and Webster, 2014) and time sampling (Lara et al., 2018a) approaches are adopted.

The area considered is divided into regions that have similar climate (e.g., wind speed and solar incidence over time), and load demand profiles. It is assumed that the potential locations for the generators and storage units are the midpoints of each region r. Additionally, generators and storage units that have the same characteristics, such as technology and operating status (i.e., existing or potential), are aggregated into clusters i and j, respectively, for each region r (Palmintier and Webster, 2014). The major impact of this approximation in the model formulation is that the discrete variables associated with generators and storage units correspond to integer rather than binary variables to represent the number of generators/storage units under a specific status in cluster i and j, respectively.

We use the same representative days as Lara et al. (2018a), selected from historical data via k-means clustering approach, where the goal of the clustering procedure is to select representative days to approximate: (i) the "duration curves" of historical load and renewables time series, (ii) the temporal correlation of each time series, and (iii) the hourly correlation between each time series. The operating scenarios are drawn from this larger set of representative days.

In order to further simplify the transmission model, the "truck-route" representation is adopted, which assumes that the flow in each line can be determined by an energy balance between nodes. This approximation ignores Kirchhoff's voltage law, which dictates that the power will flow along the path of least impedance. We also assume that the transmission lines have a maximum capacity, and that transmission expansion is not considered. Additionally, the transmission losses are characterized by a fraction loss per mile, and are not endogeneously calculated.

6.1.2 MSIP model

Assuming the data for this process is uncertain and evolves according to a stochastic process, the MSIP model can be concisely formulated as in (6.1), following the similar notation as Zou et al. (2018b).

$$\begin{aligned}\min_{(x_1,y_1)\in\chi_1} \Bigg\{ f_1(x_1,y_1) + \mathbb{E}_{\bar{\xi}_{[2,\Gamma]}|\xi_{[1,1]}} \Bigg[\min_{(x_2,y_2)\in\chi_2(x_1,\xi_2)} \Bigg\{ f_2(x_2,y_2,\xi_2) + \\ \dots + \mathbb{E}_{\bar{\xi}_{[\Gamma,\Gamma]}|\xi_{[1,\Gamma-1]}} \Bigg[\min_{(x_\Gamma,y_\Gamma)\in\chi_\Gamma(x_{\Gamma-1},\xi_\Gamma)} \Big\{ f_\Gamma(x_\Gamma,y_\Gamma,\xi_\Gamma) \Big\} \Bigg] \Bigg\} \Bigg] \Bigg\}\end{aligned} \tag{6.1}$$

where $\gamma \in \{1, ..., \Gamma\}$ is the set of stages and Γ is the last stage; x_γ is the set of state variables that link different stages; y_γ is the set of local variables that do not depend on the decision of previous stages and is only contained in the subproblem at stage γ. In the context of this GEP problem, the state variables are the number of generators in cluster i of region r that are operational in year t and number of storage units in cluster j of region r that are operational in year t. State variables are mixed-integer as the number of active generators is forced to be integer for thermal units but it is allowed to be fractional for renewable generators and storage units (for details see Appendix C.1). It is important to highlight that while decision stages are ordered in time, individual stages can have one or multiple time-periods t within it.

$\chi_\gamma(x_{\gamma-1}, \xi_\gamma)$ is the feasible region of the stage γ, which depends on the decisions in stage $\gamma - 1$ and the uncertainty realization ξ_γ in stage γ. $\bar{\xi}[\gamma, \gamma']$ denotes a sequence of random data vectors corresponding to stages γ through γ' and $\xi[1, \gamma-1]$ denotes a specific realization of this sequence of random vectors from stage 1 to stage $\gamma - 1$. $\mathbb{E}_{\bar{\xi}[\gamma,\Gamma]|\xi[1,\gamma-1]}$ denotes the expectation operation in stage γ with respect to the conditional distribution of $\bar{\xi}[\gamma, \Gamma]$ given realization $\xi[1, \gamma - 1]$ in stage $\gamma - 1$.

This stochastic process has a finite number of realizations in the form of a scenario tree $\mathcal{T}$, with Γ stages and a set of nodes in each stage denoted by S_γ. Each node n in stage $\gamma > 1$ has a unique parent node $P(n)$ in stage $\gamma - 1$. The stage containing node n is denoted by $\gamma(n)$. The set of children nodes of a node n is denoted by $C(n)$, such that if $n \in S_\gamma$ and $m \in C(n)$, then $m \in S_{\gamma+1}$. The set of nodes on the unique path from origin node 1 to node n, including the latter, is denoted by $Path(n)$. A node $n \in S_\gamma$ represents a state of the system in stage γ and corresponds to the sequence of realizations $\{\xi_m\}_{m \in Path(n)}$. The probability of node n to happen, which is the probability of realization of the sequence $\{\xi_m\}_{m \in Path(n)}$, is denoted $prob_n$. For a node in the last stage of the tree, $n \in S_\Gamma$, the sequence of realizations $\{\xi_m\}_{m \in Path(n)}$ is called a scenario $sc \in SC$ and the set of nodes n that are part of this scenario sc are denoted by S_{sc}. Therefore, the *extensive form* (also known as deterministic equivalent) of (6.1) can be formulated as:

$$\mathcal{P} : \min_{(x_n, y_n)} \left\{ \sum_{n \in \mathcal{T}} prob_n \cdot f_n(x_n, y_n) \mid (x_{P(n)}, x_n.y_n) \in \chi_n \; \forall \; n \in \mathcal{T} \right\} \tag{6.2}$$

A summary of the notation of main parts of the scenario tree is shown in Figure 6.1. The detailed MSIP formulation is described in Appendix C.1.

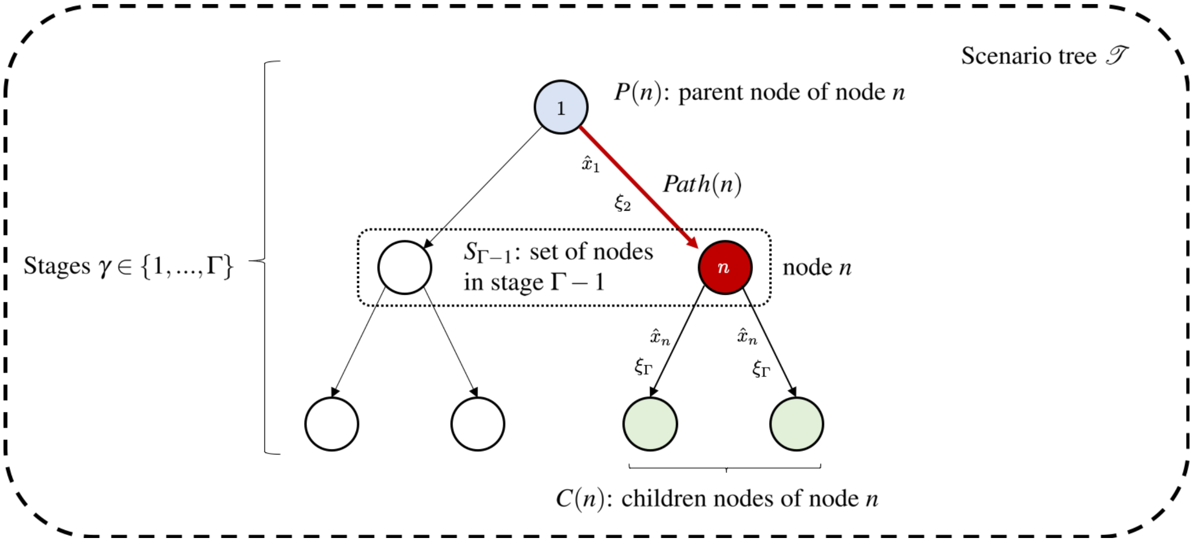

Figure 6.1: Summary of the notation for scenario tree $\mathcal{T}$

6.1.3 Scenario tree generation

In our framework we capture both long-scale strategic and short-scale operating uncertainties. The strategic uncertainties occur in the same time scale as the investment decisions (i.e., yearly), while the operating uncertainties are captured through different representative days' profiles of hourly load demand and renewable availability between investment decisions.

Our GEP problem can be represented with the multi-horizon framework proposed by Kaut et al. (2014) and used by Liu et al. (2018). This methodology represents strategic and operating uncertainties separately based on the observation that strategic decisions typically do not depend directly on any particular operational scenario, implying that it is enough to branch only between strategic stages, and the operational decisions can be seen as embedded into (or attached to) their respective strategic nodes. However, multi-horizon representation is not needed as we assume stage-wise independence in the scenario tree (see Section 6.2.2 for more details).

For a problem with Ξ^{s} strategic realizations per stage and Ξ^{o} operational realizations per stage leads to a problem with a total number of scenarios of $\overline{SC} = (\Xi^{\mathrm{s}}\Xi^{\mathrm{o}})^{\Gamma-1}$. Figure 6.2 shows the standard and the recombining representations of the scenario tree $\mathcal{T}$. They are equivalent in this context of stage-wise independence.

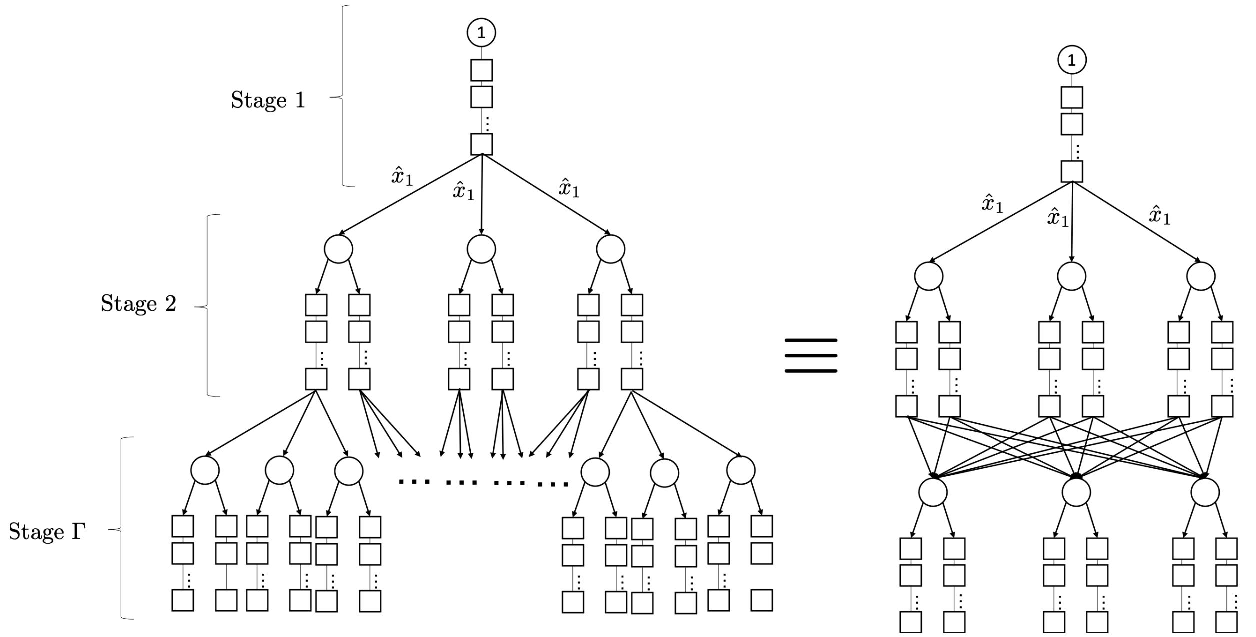

Figure 6.2: Standard (left) and recombining (right) representations of the scenario tree $\mathcal{T}$ with both strategic and operating uncertainties. The circles represent the strategic decisions and while the squares represent the operating decisions.

This scenario tree has $\Xi^{\mathrm{s}} = 3$ strategic realizations per stage and $\Xi^{\mathrm{o}} = 2$ operational realizations per stage, hence, it has a total of $\overline{SC} = (6)^{\Gamma-1}$ scenarios.

6.2 SDDiP Decomposition

As mentioned in chapter 1.2.2, multistage stochastic programming models grow exponentially with the number of stages, leading to a large multi-scale problem that quickly becomes intractable. Formulation 6.1 exploits the nested structure in this MSIP problem. Therefore, we use Stochastic Dual Dynamic Integer Programming (SDDiP) because it can take advantage this nested structure in the problem.

Birge (1985) was the first to apply Benders Decomposition on a nested fashion to solve Multistage Stochastic Linear Programming (MSLP) models. A few years later, Pereira and Pinto (1991) re-explained Nested Benders Decomposition using dynamic programming notation, incorporated scenario sampling into the algorithm, and showed that convergence can be significantly improved under the assumption of stage-wise independence, calling this new decomposition method Stochastic Dual Dynamic Programming (SDDP). Both Nested Benders Decomposition and SDDP convergence (Birge, 1985) and almost-sure convergence (Philpott and Guan, 2008) proofs, respectively, rely on the fact that cost-to-go functions in MSLP models are piece-wise linear and convex. Therefore, both decomposition methods had their application limited to convex multistage stochastic programming models (and were mostly used for linear programming models).

Recently, there have been research efforts on extending Nested Benders/SDDP to integer and mixed-integer stochastic programming models. Cerisola et al. (2009) propose a variant of Benders Decomposition for multistage stochastic integer programming and apply it to the stochastic unit commitment problem. Thome et al. (2013) introduce an extension of the SDDP framework by using Lagrangean Relaxation to convexify the recourse function applied to nonconvex hydrothermal operation planning. Zou et al. (2018b) propose a valid Stochastic Dual Dynamic Integer Programming (SDDiP) algorithm for MSIP with binary state variables, and prove that for some of the cuts presented the algorithm converges in a finite number of steps. Lara et al. (2018a) show that the cuts presented by Zou et al. (2018b) are still valid for problems with mixed-integer state variables. However, finite convergence is not guaranteed, i.e., there may be a duality gap.

The SDDiP algorithm consists of breaking down the scenario tree by nodes, and solving it iteratively in a forward and backward fashion until the optimality tolerance ϵ is satisfied, as shown in Figure 6.3. The Forward Pass yields a statistical upper bound, while the Backward Pass, which generates cuts from the relaxed subproblems to outer approximate the cost-to-go function, provides a lower bound. New cuts are added in the Backward Pass of each iteration k, and are kept in the following Forward Pass, until the difference between the upper and lower bounds is less then a pre-specified tolerance.

To be able to decompose the MSIP problem by node, we make copies of the state variables x_n. These new auxiliary variables, z_n, are used to equate to the parent node's state and make sure that when solving node n we are continuing from the state of the system at the end of its parent node. z_n is, however, relaxed to be a continuous variable within the same bounds

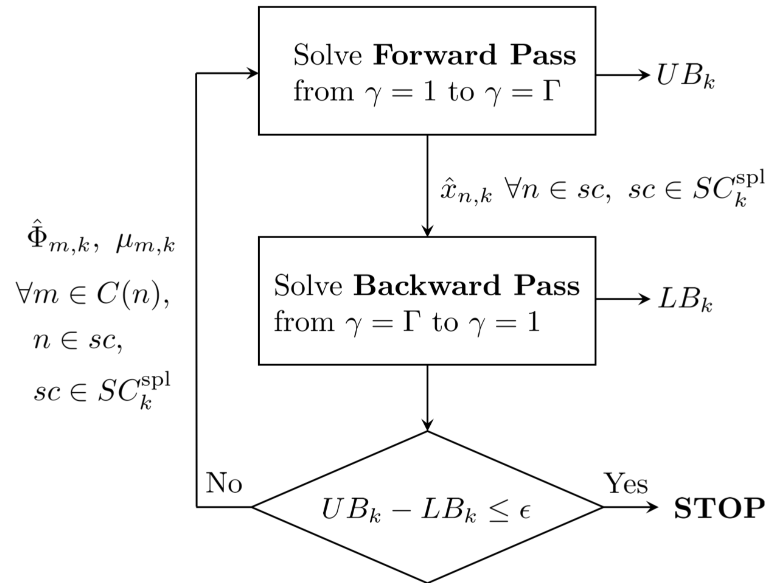

Figure 6.3: Steps at iteration k of the SDDiP algorithm, where $\hat{\Phi}_{m,k}$ and $\mu_{m,k}$ are the coefficients of the cuts, sc is a scenario and SC^{spl_k} is the set of scenarios sampled in iteration k.

as x_n.

Going back to our MSIP problem, the node subproblems of (6.2) can be formulated as follows $\forall\ n \in \mathcal{T}, k \in \mathcal{K}$:

$$
\begin{aligned}
\mathcal{P}_{n,k} : \Phi_{n,k}(\hat{x}_{P(n),k}, \phi_{n,k}) := \quad & \min_{(x_n,y_n)} \quad f_n(x_n, y_n) + \sum_{m \in C(n)} q_{nm} \phi_{m,k}(\hat{x}_{n,k}) \\
& s.t. \quad (z_n, x_n . y_n) \in \chi_n \\
& \qquad\quad z_n = \hat{x}_{P(n),k} \quad \leftarrow \mu_{n,k} \in \mathbb{R}^{\ell} \\
& \qquad\quad x_n \in \mathbb{Z}_+^{\ell_1} \times \mathbb{R}_+^{\ell_2}, \quad y_n \in \mathbb{Z}_+^{o_1} \times \mathbb{R}_+^{o_2}, \quad z_n \in \mathbb{R}^{\ell}
\end{aligned}
\tag{6.3}
$$

where $\ell = \ell_1 + \ell_2$, $o = o_1 + o_2$; $\mathcal{K}$ is the set of iterations; and $q_{nm} := prob_m/prob_n$ is the conditional probability of transitioning from node n to node m for $m \in \mathcal{T} \setminus \{1\}$ and $n = P(m)$.

The approximate expected cost-to-go function, $\phi_{n,k}(\cdot)$, is defined as:

$$
\phi_{n,k}(\hat{x}_{n,k}) := \min_{x_n,\alpha_n} \left\{ \alpha_n :\ \alpha_n \geq \sum_{m \in C(n)} q_{nm} \cdot \left(\hat{\Phi}_{m,k'} + \mu_{m,k'}^{\intercal} (\hat{x}_{n,k'} - x_n) \right)\ \forall\ k' \in \mathcal{K} | k' < k \right\}
$$

(6.4)

6.2.1 Forward Step

The Forward Step for the SDDiP is very similar to the Forward step presented by Lara et al. (2018a), but now applied to the scenario tree with incorporated scenario sampling.

As mentioned before, the purpose of the Forward Pass in SDDiP is to generate a statistical upper bound to the solution of the MSIP for the entire scenario tree $\mathcal{T}$. It accomplishes this by randomly sampling a subset of the scenarios in the tree $\mathcal{T}$ - SC_k^{spl}, and for each stage $\gamma \in \{1, ..., \Gamma\}$ solving the nodes n in $S(\gamma)$ if they are also part of the sampled scenarios. By solving node n, the algorithm implements the optimal decisions for this node considering its uncertainty realization and previous $Path(n)$, and passes the current state of the system forward to its children node $m \in C(n)$ if m is also part of the sampled scenarios. This process is repeated up until all sampled scenarios are fully solved (until last stage Γ) and we have a total minimum cost for each of those scenarios.

The Forward Pass with scenario sampling is shown in Figure 6.4 for both the standard representation and the recombining scenario tree representation. For our case in which we assume stage-wise independence, both representations are equivalent.

The problem is assumed to have complete continuous recourse, which means that for any value of state variable (i.e., linking variable) and local integer variables, there are values for the continuous local variables such that the solution is feasible. This assumption is valid since feasibility can be achieved by adding nonnegative slack variables and penalizing them in the objective function.

The statistical upper bound, UB_k, is calculated in the Forward Pass as in (6.5).

$$UB_k = \bar{\mu}_k + z_{\alpha/2} \cdot \frac{\sigma_k}{\sqrt{N^{\text{spl}}}} \qquad \forall\ k \tag{6.5}$$

where $\bar{\mu}_k$ is the mean total cost over the sampled scenarios in iteration k, σ_k is its standard deviation, N^{spl} is the number of scenarios sampled in each iteration, and $z_{\alpha/2}$ is the z-score to

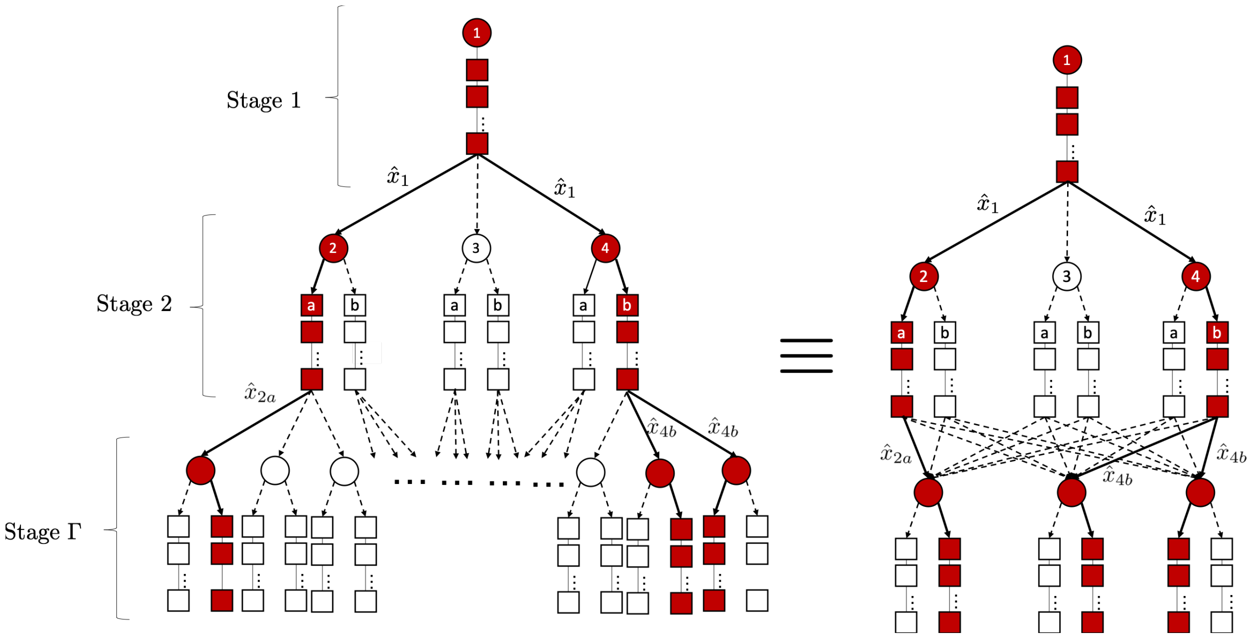

Figure 6.4: Forward Pass with scenario sampling for both the standard and the recombining scenario tree representations. The nodes highlighted are the ones that are part of the sampled scenarios, and the continuous arrows show the paths of the sampled scenarios.

assure a certain confidence interval (for example, for a 95% confidence interval, $z_{\alpha/2} = 1.96$).

The total cost of a scenario sc is as follows.

$$\Phi^{\text{tot}}_{sc,k} = \sum_{\gamma \in \{1,\dots,\Gamma\}} \sum_{n \in S_\gamma \,\cap\, n \in S_{sc}} \left(\hat{\Phi}_{n,k} - \hat{\alpha}_n \right) \qquad \forall sc \in SC^{\text{spl}}_k \tag{6.6}$$

and $\bar{\mu}_k$ and σ_k are defined in (6.7) and (6.8), respectively.

$$\bar{\mu}_k = \frac{1}{N^{\text{spl}}} \sum_{sc \in SC^{\text{spl}}_k} \Phi_{sc,k} \qquad \forall k \in \mathcal{K} \tag{6.7}$$

$$(\sigma_k)^2 = \frac{1}{N^{\text{spl}} - 1} \sum_{sc \in SC^{\text{spl}}_k} (\Phi_{sc,k} - \bar{\mu}_k)^2 \qquad \forall k \in \mathcal{K} \tag{6.8}$$

It is important to note that the upper bound, UB_k, obtained by the SDDiP is only a statistical upper bound. Its validity is guaranteed with certain probability provided that N^{spl} is not too small. However, regardless of the size of N^{spl}, it is possible that the upper bound is smaller than the valid lower bound evaluated in the backward step. To avoid this issue, alternative stopping criteria are reported in the literature (Homem-de Mello et al., 2011; Shapiro et al., 2013a; Bruno et al., 2016). We, however, are using the standard stopping criteria of $UB_k - LB_k \leq \epsilon$, where ϵ is the allowed optimality tolerance.

6.2.2 Backward Step

After solving the Forward Pass, the next step is the Backward Pass, and its purpose is to generate cuts that outer approximate the cost-to-go function. The Backward Pass consists of solving the subproblems from the last to the first stage, so the solutions of future stages can be used to generate cuts and provide approximations to the cost-to-go functions within the planning horizon.

Since our MSIP has mixed-integer recourse (i.e., the state variables x_n are mixed-integer), instead of solving the original subproblems we have to solve a relaxation that is convex in the subspace of the state variables in order to generate a valid cut. This relaxation can be the linear programming relaxation or the Lagrangean relaxation of the subproblem $\mathcal{P}_{n,k}$ given by (6.3). Depending on the type of relaxation solved, a different type of cut is added in the Backward Pass to approximate the cost-to-go function.

In this step, instead of only having to solve the nodes that are part of the scenarios sampled in iteration k, SC_k^{spl}, we have to solve the subproblem of all the children nodes $C(n)$ of the nodes n that are part of the sampled scenarios $n \in S_{sc}, sc \in SC_k^{\text{spl}}$. Consequently, we solve a total of $(N^{\text{spl}} \cdot \Xi^{\text{s}} \cdot \Xi^{\text{o}})$ subproblems, as can be seen in Figure 6.5. The solution of this extra set of nodes is necessary to be able to generate the cuts that approximate the cost-to-go function, which takes an weighted average of the coefficients coming from the solution of the subproblem of the children nodes based on the probabilities of the uncertainty realizations.

The lower bound, LB_k, is calculated in the Backward Pass as in (6.9). It is easy to see that the relaxed solution of the root node $n = 1$ is a lower bound to the total cost since it only has a subset of the original constraints of the original problem.

$$LB_k = \hat{\Phi}_{1,k} \qquad \forall\ k \in \mathcal{K} \tag{6.9}$$

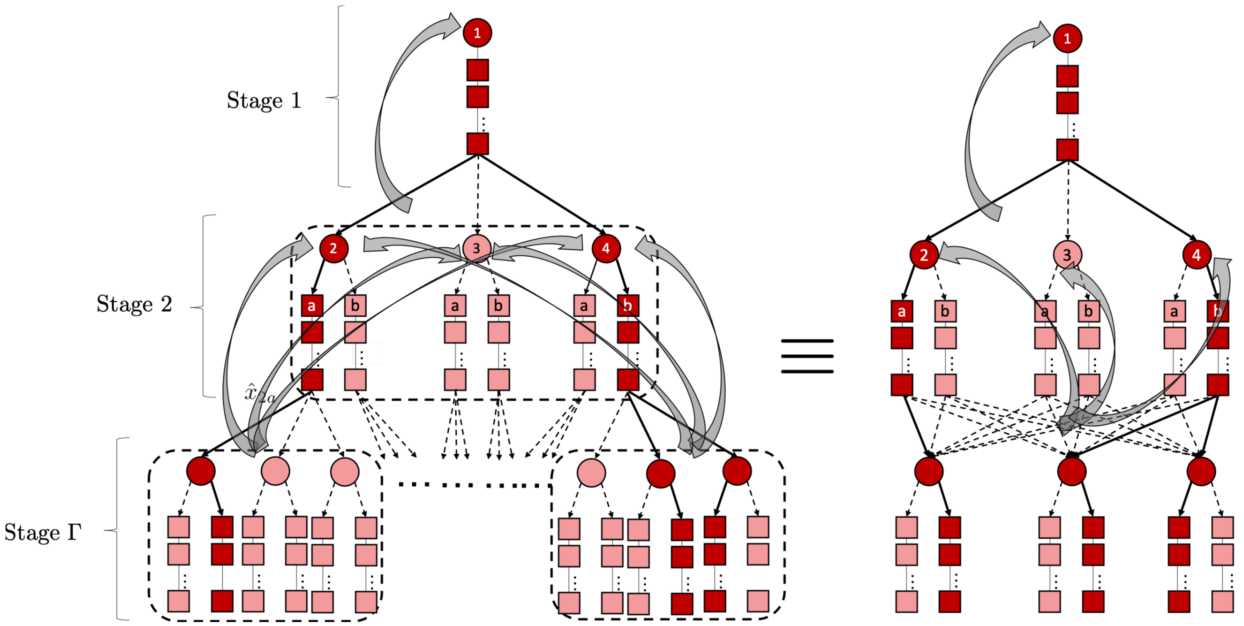

Figure 6.5: Backward Pass with scenario sampling and stage-wise independence for the standard (left) and recombining (right) scenario tree representations. The nodes in dark color are the ones that are part of the sampled scenarios, and the nodes in lighter color are the nodes that are not part of the sampled scenarios but were solved because they are children nodes of the sampled nodes.

Possible cuts to approximate the cost-to-go function

The choice of cuts directly impacts the performance of the algorithm as some cuts are tighter and more/less computationally expensive to generate than the others. The Benders cut, Strengthened Benders cut and Lagrangean cut were first proposed by Zou et al. (2018b) for MSIP with binary recourse, and Lara et al. (2018a) proved their validity for models with mixed-integer recourse. However, it is important to highlight that the SDDiP algorithm does not have guaranteed finite convergence if the formulation has mixed-integer recourse and the Backward Pass uses any of the following cuts, which means that there can be a duality gap.

Benders cut The first option trivially comes from SDDP Benders cut. The Benders cut's coefficients are obtained from the solution of the linear relaxation (LP) of $\mathcal{P}_{n,k}$ in (6.3), and is formulated as follows.

$$\alpha_n \geq \sum_{m \in C(n)} q_{nm} \cdot \Big(\hat{\Phi}^{\mathrm{LP}}_{m,k'} + {\mu^{\mathrm{LP}}_{m,k'}}^{\intercal} (\hat{x}_{n,k'} - x_n) \Big) \qquad \forall\, k' \in \mathcal{K} | k' < k \tag{6.10}$$

This is the weakest of the possible cuts, but it has the advantage of being easily and quickly computed. As shown by Lara et al. (2018a), the Benders cut performs very well for this GEP problem, since it has a tight linear relaxation. For certain multistage capacity planning problems with integer recourse, there is evidence that Benders cuts alone are sufficient for reducing the optimality gap to zero as the number of stages increases (see, e.g., Huang and Ahmed (2009), Corollaries 1 and 2).

Lagrangean cut The subproblem $\mathcal{P}_{n,k}$ can also be convexified by considering its Lagrangean relaxation, which yields the convex hull of the nonlinking constraints (Frangioni, 2005). In our case this is done by dualizing the linking equalities in (6.3) and penalizing their violation in the objective function by the vector of Lagrange multipliers, $\mu_{n,k}$. The closer the Lagrange multipliers are to their optimal value, the tighter the approximation is, and the stronger the cuts generated by theses multipliers are. Therefore, the Lagrangean cut uses the coefficients obtained by the Lagrangean Dual (LD) problem and is formulated as follows.

$$\alpha_n \geq \sum_{m \in C(n)} q_{nm} \cdot \Big(\hat{\Phi}^{\mathrm{LD}}_{m,k'} + {\mu^{\mathrm{LD}}_{m,k'}}^{\intercal} (\hat{x}_{n,k'} - x_n) \Big) \qquad \forall\, k' \in \mathcal{K} | k' < k \tag{6.11}$$

For more details on the Lagrangean cut, see Zou et al. (2018b); Lara et al. (2018a).

Strengthened Benders cut In order to mitigate potential performance issues, Zou et al. (2018b) proposed the Strengthened Benders cut, which is a compromise between Benders and Lagrangean cuts. Its generation is similar to the Lagrangean cut, but it does not use the subgradient method to improve the multipliers. Instead, it uses the coefficients from the first Lagrangean relaxation (LR) solved after the initialization of the multipliers using LP relaxation and is formulated as follows.

$$\alpha_n \geq \sum_{m \in C(n)} q_{nm} \cdot \left(\hat{\Phi}^{\text{LR}}_{m,k'} + {\mu^{\text{LP}}_{m,k'}}^{\intercal} (\hat{x}_{n,k'} - x_n) \right) \qquad \forall\, k' \in \mathcal{K} | k' < k \qquad (6.12)$$

For more details on the Strengthened Benders cut, see Zou et al. (2018b); Lara et al. (2018a).

Due to the computational expense of computing the Lagrangean and the Strengthened Benders cuts, and the computational evidence in Lara et al. (2018a) that the Benders cuts are likely sufficient for this GEP problem, we select the Benders cuts to be used in our computational experiments shown in Section 6.3.

Stage-wise independence and Cut Sharing

In multistage problems, if the stochastic process and the constructed scenario tree is stage-wise independent, i.e., for any two nodes n and n' in S_t the set of children nodes $C(n)$ and $C(n')$ are defined by identical data and conditional probabilities, then the cost-to-go functions do not depend on the current scenario. This means that the value functions and expected cost-to-go functions depend only on the stage rather than the nodes, $\Phi_n(\cdot) \equiv \Phi_\gamma(\cdot)\ \forall\, n \in S_\gamma$, and the cuts generated for a particular scenario are also valid for any other scenario at the same stage (Infanger and Morton, 1996).

The SDDiP relies on the stage-wise independence assumption and the ability to share cuts among different nodes in the same stage to avoid the combinatorial explosion and the "curse of dimensionality" (Pereira and Pinto, 1991). Therefore, it provides a practical solution for solving real-world applications of MSIP on very large scenario trees without the need for scenario reduction methods.

The Backward Pass in the SDDiP algorithm works similarly as the one in a Stochastic Nested Benders decomposition with sampling. The only difference is that because we have stage-wise independence, the cuts generated are added to all the nodes in the previous stage instead of only to the parent node. The Backward Pass with stage-wise independence assumption is shown is Figure 6.5.

The stagewise independence assumption is reasonable for the operating uncertainties considered in our GEP problem (i.e., different profiles for representative days) since the realization of the solar incidence, wind speed and load profile in one day has little influence in the next day, especially if we are only including a few representative days a year. Regarding the strategic uncertainties, the stage-wise independence assumption is adequate for natural gas price uncertainty as the prices can go up and down without a clear influence of the realization in the year before. In the case of carbon tax uncertainty, this assumption may be a stretch as it is unlikely that the carbon tax would wildly vary between two consecutive years. However, even though carbon tax stage-wise independence generates an exceptionally uncertain mode, we still believe that in face of the the political uncertainty and changes in administration over the planning horizon this is a case worth examining.

Additionally, there has been some work on how to extend the use of SDDP/SDDiP for certain types of interstage dependency (Infanger and Morton, 1996; Shapiro et al., 2013a; de Queiroz and Morton, 2013; Lohmann et al., 2016), which would be useful to capture GEP uncertainties that are not well-captured by the stage-wise independence assumptions (e.g., learning rate of new generation and storage technologies and peak load). For the case of strategic uncertainty with stage-wise dependence and operating uncertainty with stage-wise independence, one can use the framework proposed by Rebennack (2016), which combines SDDP with the sampling-based stochastic nested Benders decomposition approach. These extensions are left for fuure investigation.

6.2.3 Parallelization Scheme

In the SDDiP framework the subproblems of the nodes n within the stage γ, $n \in S_\gamma$, are independent from each other. Hence, SDDiP is well suited for parallel processing. The algorithm is not, however, trivially parallel since synchronization is required to share the Benders cuts generated with all the nodes in the previous stage $\gamma - 1$. Therefore, a well-thought parallelization strategy is required in order to obtain an efficient parallel solution.

There has been some effort in the literature to propose the optimal parallelization scheme for SDDP in order to avoid synchronization steps as much as possible (Pinto et al., 2013; Helseth and Braaten, 2015). We do not claim that our parallelization scheme is optimal, but based on the results that we obtained it seems to be adequate for our problem.

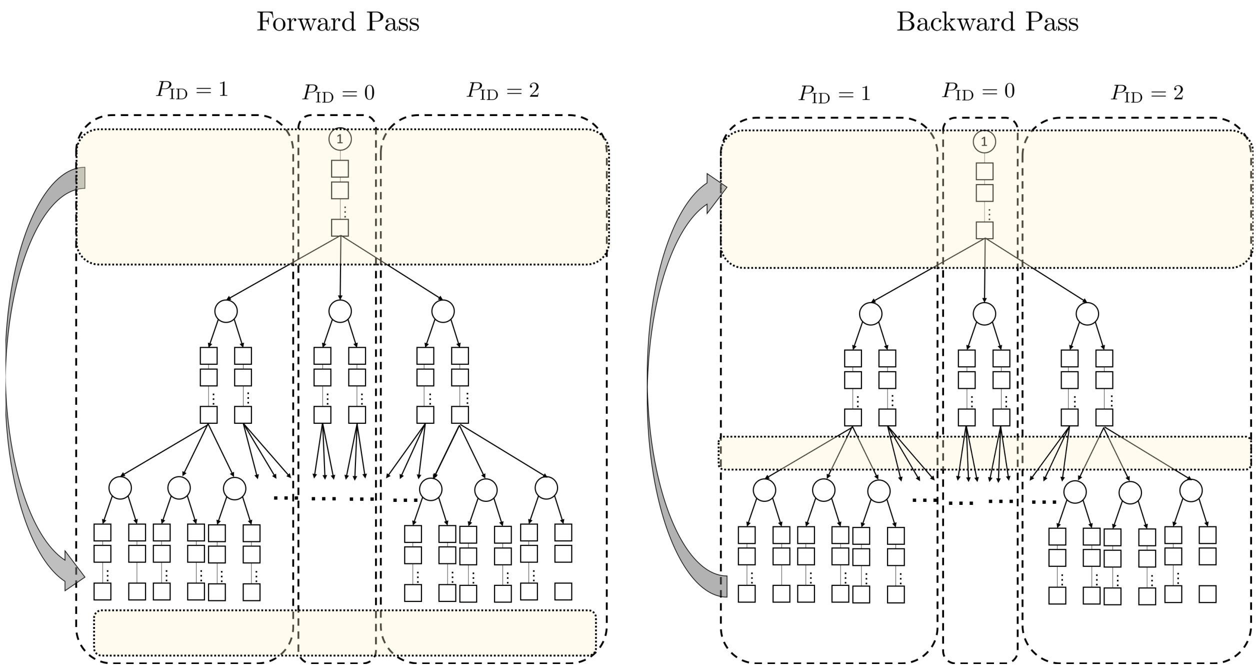

Figure 6.6: Parallelization scheme used for the SDDiP algorithm. The dashed lines show which process is in charge of which node, and the highlighted areas show the synchronization points both in the Forward and Backward Passes.

We use PyMP (Lassner, 2018) for the parallelization, which is a Python package based on OpenMP (Chandra et al., 2001). We first equally divide the nodes in the tree among the processes, and then enter the parallel context and have each process generate the subproblems for those nodes assigned to it. After that, we start the Forward Pass and randomly select the sampled scenarios to be solved for each of those processes, $sc \in SC_k^{\mathrm{spl,pid}}$, such that the number of processes N^{pids} times the number of sampled scenarios by processes $N^{\mathrm{spl.pid}}$ equals the total number of sampled scenarios per iteration k: $N^{\mathrm{pids}} \cdot N^{\mathrm{spl,pid}} = N^{\mathrm{spl}}$. We then solve the subproblems of all the nodes that are part of the sampled scenarios, storing the results of the state variables as shared dictionaries among the processes. Note that as the nodes are statistically assigned to processes there is the potential for load imbalance due to both the random sample of scenarios and because of variance in the time to solve each MILP subproblem. Addressing this potential scaling issue is left for future work.

When all processes reach the end of the Forward Step, there is a synchronization step to gather all the optimal values before the the first process calculates the upper bound UB_k and distributes it to the other processes. In the Backward Pass there are synchronization steps after every stage to distribute cuts generated at that stage to all nodes in the previous stage. In the end only the first process calculates the lower bound LB_k and check if the optimality tolerance ϵ is satisfied. The parallelization scheme and the synchronization points for both the Forward and Backward Passes are shown in Figure 6.6.

Our implementation of the parallel SDDiP algorithm for this GEP problem can be found in Lara (2019).

6.3 Case study: ERCOT region

We test the proposed MSIP formulation and SDDiP algorithm for a case study approximating the Texas Interconnection, a power grid that covers most of the state of Texas and is managed by the Electric Reliability Council of Texas (ERCOT). This case study is based on the deterministic case study presented by Lara et al. (2018a), with the addition of operational and strategic uncertainties, and the option of adding storage units.

Within the ERCOT covered area, we consider four geographical regions: Northeast, West, Coastal and South. We also include a fifth region, Panhandle, which is technically outside the ERCOT limits but due to its renewable generation potential, it supplies electricity to the ERCOT regions. Thus, Panhandle is considered a zone with zero load demand, i.e., it is only a supplier, not a consumer. The regions are shown in Figure 5.1. For more details on the sources of the data used, see Lara et al. (2018a).

We consider 3 types of utility batteries: lithium-ion, lead-acid and flow batteries, for which we use the capital cost forecast provided by Schmidt et al. (2017) and the technical information provided by Luo et al. (2015), the same sources used by Lara et al. (2018b).

6.3.1 Reference case: all energy sources included

For each of the regions, we use load and capacity factor profiles with an hourly resolution. Representative days are constructed using a k-means clustering algorithm and 2004-2010 zonal load and renewables profiles, as explained in Lara et al. (2018a). 8 clusters are constructed to find the 8 most representative days, and these days are split in 2 scenarios which correspond to 2 realizations of the operational uncertainty per stage, with 4 representative days each. We can assume that the operational uncertainty satisfies the stage-wise independence assumption discussed in Section 6.2.2 and follows a uniform distribution.

Additionally, we consider that the natural gas fuel price is uncertain and has 3 realizations per stage. We assume that this fuel price is stage-wise independent and follows a uniform distribution. The realizations were built using the minimum, median and maximum value corresponding to the scenarios presented in the EIA Annual Energy Outlook 2019 (EIA, 2019). We assume that the coal and uranium prices are deterministic since they exhibit considerably less variation compared to the natural gas price.

Our computational tests are performed on a MacBook Pro with 2.3GHz quad-core 8th-generation Intel Core i5 processor, with 8GB of RAM, running on macOS 10.14 Mojave. We implement the SDDiP algorithm in Python 3.6.6 and Pyomo 5.6.1, and solve the LPs and MILPs of each node of the the scenario tree using Gurobi version 8.0.1 (Gurobi Optimization, 2018). We allow a total number of 3 parallel processes, sample 15 scenarios per iteration, impose a 95% confidence interval in the statistical upper bound ($z_{\alpha/2} = 1.96$), and consider that the algorithm converges if it reaches an optimality gap of less than or equal to 1%.

We first test the parallel efficiency of our algorithm by solving the same 5-stage problem both sequentially and in parallel (we fix the random seed to avoid the stochasticity in the solution by the random sampling of scenarios per iteration). The solution time for the serial implementation is $t_s = 13,950$ seconds and for the parallel implementation with 3 processes is $t_p = 6,131$. Therefore, our parallel SDDiP algorithm has 76% efficiency.

To test the capabilities of our algorithm, we solve the problem for: (i) 5 stages (5 years, 1 year per stage); (ii) 10 stages (10 years, 1 year per stage); and (iii) 15 stages (15 years, 1 year per stage). In all of them, the subproblem (node) size before cuts is 50,042 constraints, 13,746 integer variables, and 22,755 continuous variables. The size of the extensive form (deterministic equivalent) and performance of the SDDiP algorithm for all cases are reported in Table 6.1.

Table 6.1: Size of the problem and SDDiP performance for scenario trees with different numbers of stages

	5 stages	10 stages	15 stages
Number of scenarios	$1,296$	1.01×10^7	7.84×10^{10}
Number of nodes in the scenario tree	$1,555$	1.21×10^7	9.40×10^{10}
Number of constraints (extensive form)	7.78×10^7	6.05×10^{11}	4.71×10^{15}
Number of integer variables (extensive form)	2.14×10^7	1.66×10^{11}	1.29×10^{15}
Number of continuous variables (extensive form)	3.54×10^7	2.75×10^{11}	2.14×10^{15}
Wall-clock time [s]	$6,131$	$8,146$	$86,049$
Upper bound [$ billion]	51.69	91.87	122.40
Lower bound [$ billion]	51.69	91.27	121.72
Optimality gap [%]	8.57×10^{-4}	0.66	0.56

The extensive forms of all three cases are massive, with up to quadrillions of variables and constraints in the 15-stage case. Considering the size of the models, all cases were solved in a reasonable amounts of time: 1.7 hours, 2.3 hours and 23.9 hours, respectively. Due to memory limitations, it is fair to say that at least the 10-stage and 15-stage cases would not be solvable in a personal laptop or desktop without decomposing the model, or if the model is decomposed by a scenario-based approach without scenario reduction techniques. Additionally, the convergence time is considerably less than similar sized GEP problems with multistage stochastic programming formulations reported in the literature that use scenario-based decomposition (see Liu et al. (2018)). These results show how powerful and useful SDDiP can be for practical large-scale MSIP models.

An important question that can now be explored is how impactful the length of the planning horizon and the number of stages are in the optimal "here-and-now" first stage investment decisions. For this reference case, even though there are marginal differences between the optimal first-stage decisions depending on the number of stages (e.g. in the 15-stage solution the optimization adds 2 PV solar units in the first year while the 5-stage and 10-stage solutions do not), the results are extremely similar (2.27 GW, 2.21 GW, 2.40 GW of natural gas generation capacity are added in year one, respectively), indicating that solving a 5-stage problem would be sufficient.

This result is not surprising considering that the ERCOT system is mature and stable, and therefore we would not expect to see drastic changes without a corresponding drastic impulse on the system. Accordingly, the optimization chooses to install as few generators as possible in this first year, and wait until some of the uncertainty is realized to make future decisions.

6.3.2 *No nuclear* case

The value of stochastic programming with multiple decision stages becomes more accentuated in power systems where there is a need for quick and significant expansion, which is the case in some developing countries (e.g., India's electricity demand is expected to double over the coming decade (IEEFA, 2017)).

Therefore, we solve a hypothetical case in which ERCOT decides to immediately retire nuclear power. This is an interesting analysis considering the declining profits and scheduled retirements of nuclear plants in the United States (Union of Concerned Scientists, 2018). The nuclear reactors represent 5% of the initial ERCOT generation capacity in the original data set. Thus, by imposing their immediate retirement, the optimization is forced to make significant expansion decisions in the first stage (first year).

We first solve the *no nuclear* case study with the same scenario tree as before, i.e. 2 realizations of operational uncertainty per stage, 3 realizations of natural gas price per stage, and 15 stages. The optimal ERCOT generation capacity by source in the first year (first stage) for the reference case (all sources included and natural gas price uncertainty) and the *no nuclear* case with natural gas price uncertainty are shown in Figure 6.7. The results show that the nuclear plants were fully replaced by natural gas combined-cycle plants.

A potential issue that would arise if all nuclear reactors are replaced by natural gas turbines is the increase of CO_2 emissions (Union of Concerned Scientists, 2018). Therefore, we

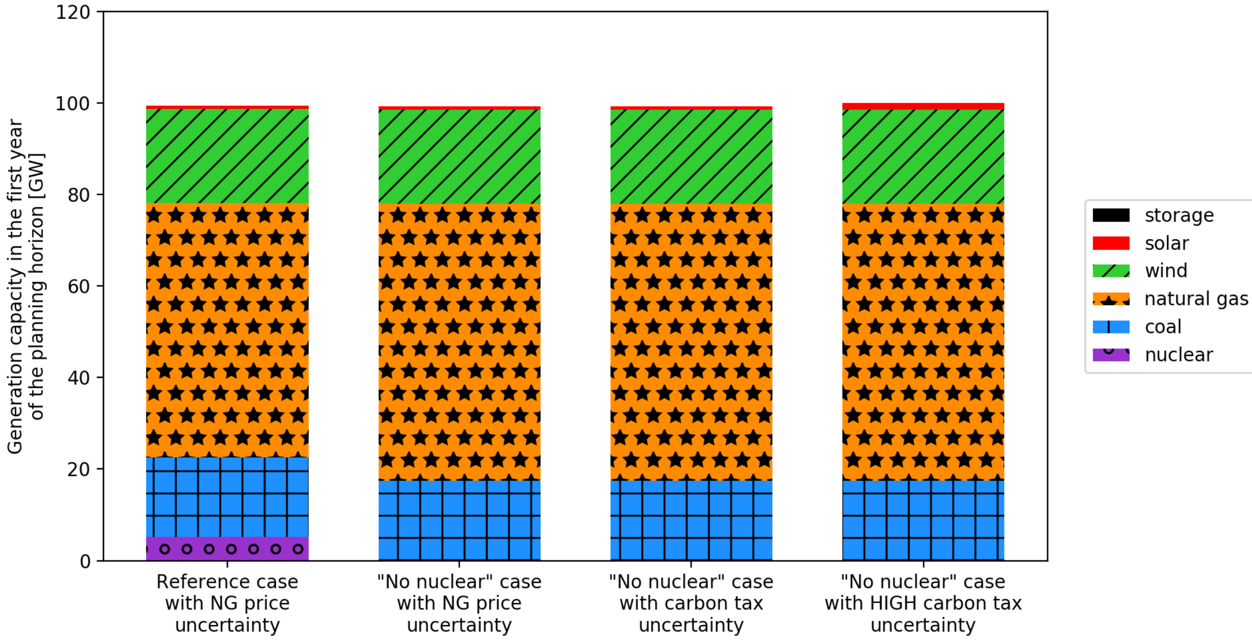

Figure 6.7: ERCOT generation capacity by source in the first year (origin node) for the reference case (all sources included and natural gas price uncertainty) and the *no nuclear* case with natural gas price uncertainty, carbon tax uncertainty, and high carbon tax uncertainty.

also solve the *no nuclear* case study with carbon tax uncertainty. We consider 2 realizations of operational uncertainty per stage (same as before), and 3 realizations of carbon tax price per stage such that the *no* realization of carbon tax equals \$0.0/tonne CO_2 for all stages, the *high* realization starts at \$10/tonne CO_2 at year 2 (stage 2) and increases linearly to \$150/tonne CO_2 at year 15 (stage 15), and the *medium* realization is the average between *low* and *high* realizations. We assume that the carbon tax prices are stage-wise independent and follow a uniform distribution. As mentioned in Section 6.2.2, the carbon tax stage-wise independence assumption generates an "exceptionally uncertain" model. The optimal ERCOT generation capacity by source in the first year (first stage) for the *no nuclear* case with carbon tax uncertainty is also shown on the right of Figure 6.7. The results show that even with the risk of having carbon tax fees in the future, the optimal "here-and-now" decision is to invest on new natural gas combined-cycle plants.

Additionally, we solve the *no nuclear* case study with a scenario tree that considers more extreme realizations of carbon tax: the *low* realization of carbon tax equals \$0.0/tonne CO_2

for all stages (same as before), the *high* realization starts at \$100/tonne CO_2 at year 2 (stage 2) and increases linearly to \$500/tonne CO_2 at year 15 (stage 15), and the *medium* realization is the average between *low* and *high* realizations. We assume again that the carbon tax prices are stage-wise independent and follow a uniform distribution.

The reasoning behind solving the *no nuclear* case study for this more extreme scenario tree is to find out if natural gas will stop being the most attractive source if the variability between the realizations of carbon tax is higher. The optimal ERCOT generation capacity by source in the first year (first stage) for the *no nuclear* case with high carbon tax uncertainty is also shown on the right of Figure 6.7. The results show that the risk of having steep carbon tax fees in the future makes the optimization invest less in natural gas technologies in the first year (reduction of 0.58 GW in natural gas generation capacity), and more in renewable sources (increase of 0.74 GW and 0.04 GW in solar and wind generation capacity, respectively).

6.3.3 The value of stochastic solution

In order to evaluate the potential gain of solving this GEP as a MSIP model instead of a deterministic model, we solve a deterministic version of the *no nuclear* case study with high carbon tax uncertainty using the average of the carbon tax realizations. The comparison between the generation capacity by generation technology in the first year (first stage) for the MSIP formulation and the deterministic formulation is shown in Figure 6.8.

The results show that by assuming that carbon tax is a deterministic parameter, the optimization makes more conservative decisions and replaces some of the natural gas combined-cycle (ng-cc) and gas-fired combustion turbine (ng-ct) by natural gas combined-cycle with carbon capture (ng-cc-ccs) to avoid having to pay carbon tax for their emissions later. It also installs more wind turbines in the deterministic case than in the stochastic case. These results make sense because in the deterministic case the optimization is sure that there will be carbon tax in the future, while in the stochastic case there may be no carbon tax, a high carbon tax or a medium carbon tax, therefore it is better to wait and get more information

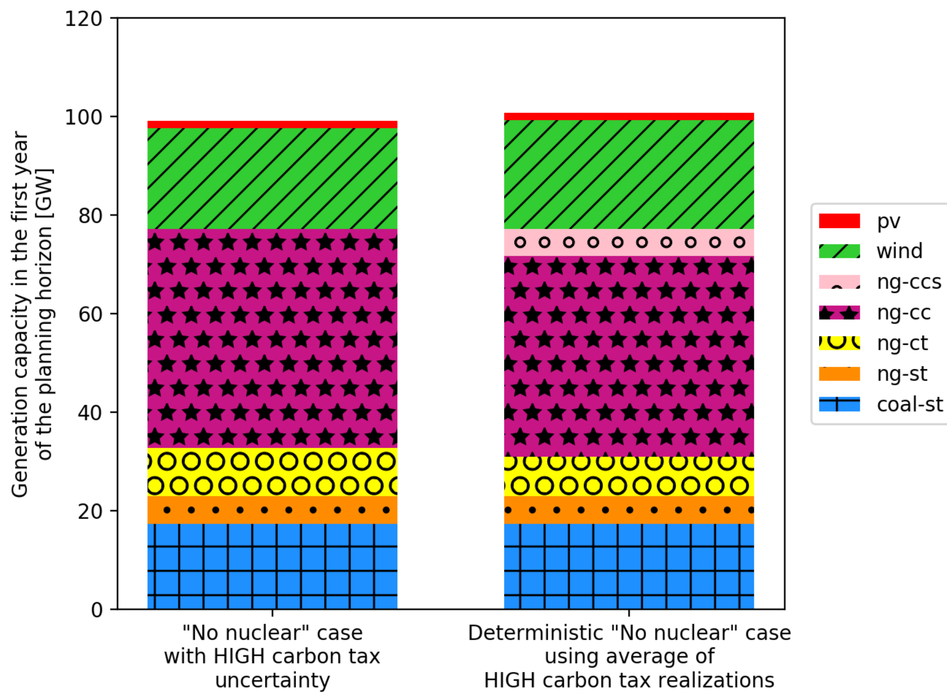

Figure 6.8: ERCOT generation capacity by generation technology in the first year (origin node) for the *no nuclear* case with high carbon tax uncertainty and the deterministic solution using the carbon tax averages.

about the carbon tax realization before investing in more expensive low-emission options.

To evaluate how well the deterministic first-stage solution would perform in our high carbon tax uncertainty scenario tree, we re-solve the MSIP formulation for the *no nuclear* case study with high carbon tax uncertainty fixing the investment decisions in the first stage to be the ones given by the deterministic solution. While the original stochastic solution gives an optimal expected value of \$ 232.47 billion (0.28% optimality gap) the stochastic solution using the first-stage solution of the deterministic model gives an expected value of \$ 234.65 billion (0.20% optimality gap), showing that by considering carbon tax an uncertain parameter the value of the stochastic solution is \$2.18 billions, which is the savings one can achieve in the long term.

6.4 Conclusion

In this chapter, we have proposed a multistage stochastic mixed-integer programming formulation to address the long-term generation expansion planning under uncertainty including operating details in the hourly level, storage options, and multiscale representation of uncertainty (strategic and operational). We decompose the model using a parallelized SDDiP algorithm and show that this is a powerful framework for solving practical large-scale MSIP formulations, allowing the solution of models with quadrillions of variables and constraints in a personal computer.

We solved a hypothetical case study for the ERCOT region and show that for most of the scenario trees tested (with natural gas price and carbon tax uncertainty) the first-stage decisions consist of investing in new natural gas plants, indicating the competitiveness of this source for this case study. We also ran a case study in which all nuclear reactors are immediately retired, and unless we consider steep values for the high realization of carbon tax the optimization still decides to invest in natural gas turbines in the first stage. Finally, we show the value of stochastic solution for the scenario tree with high carbon tax uncertainty, with a potential $ 2 billion reduction in cost in the long run.

As future work it would be interesting to consider the lead time of construction of the power plants as both a deterministic and an uncertain parameter to evaluate how this would impact in the planning strategy. Another addition to this work would be to evaluate and improve the parallel scalability of the proposed algorithm.

Chapter 7

Conclusions

7.1 Summary of this Thesis

In this section we summarize the major findings and accomplishments in each chapter.

7.1.1 Global Optimization Algorithm for Capacitated Multi-facility Continuous Location-Allocation Problems

In chapter 2 we have presented a new version of the Capacitated Multi-facility Weber Problem (CMWP) that considers: (i) fixed costs for opening new facilities, and fixed transportation costs; (ii) multiple types of facilities; and (iii) two sets of fixed points representing suppliers and consumers. CMWP models assume 2-dimensional continuous variables for the location of facilities. In principle this may not be considered to be practical and one may prefer fixed pre-specified locations. However, for large areas in which there is no prior knowledge of what could be potentially good locations, handling them as discrete might lead to an intractable problem. Therefore, one can interpret the CMWP model proposed in this chapter as a higher level screening tool which could be subsequently refined to discrete locations near those points identified by the proposed model. We have proposed the GDP formulation (2.1), which is nonconvex due to the bilinear terms in the transportation cost. The presence of nonconvexities was the main motivation for representing the problem as a

GDP. By having the bilinear terms as part of the disjunctions, the transportation costs are calculated only for the selected connections within an iterative procedure. We also assume that the facilities of the same type have the same characteristics and costs associated with them. Therefore, we have the additional constraints (2.2w)-(2.2x) to break the symmetry in the facility selection within the same type and avoid degeneracy in the solution. The GDP was reformulated into a nonconvex MINLP (2.2) using the hull reformulation.

Although global optimization solvers (BARON, ANTIGONE and SCIP) perform reasonably well for small-scale instances of the nonconvex MINLP problem (2.2), their performance scales poorly due to the loose bounds of the variables in the bilinear term, thereby becoming computationally very expensive for mid to large-scale problems. For this reason, we have proposed a bilevel decomposition algorithm that consists of decomposing the problem into a master problem and a subproblem. The master problem is based on a relaxation of the nonconvex MINLP (2.2) by discretizing the 2-dimensional space, which yields an MILP that predicts the selection of facilities and their links to suppliers and customers, as well as a lower bound on the cost of problem (2.1) or (2.2). The subproblem corresponds to a nonconvex NLP of reduced dimensionality that results from fixing the binary variables in the MINLP problem (2.2), according to the binary variables predicted in the MILP master problem. We have proved that this algorithm converges to the global optimum within an ϵ tolerance in a finite number of iterations.

We first illustrated the proposed algorithm by solving a small test problem with 2 suppliers and 2 customer points. The results show that the lower bound gradually tightens up as the number of iterations proceeds and, consequently, the number of partitions increase. The optimal solution of 5039.3 is the same as found by the general purpose global optimization solvers. However, while the algorithm solved this problem within an optimality tolerance of 0.5% in 11.9 seconds, BARON solved it in 1247.1 seconds, and ANTIGONE and SCIP could not solve it in 1 hour.

We have tested the algorithm for 15 random test cases varying from 2 suppliers and 2 consumers, to 40 suppliers and 40 consumers, from 1 to 3 types of facilities, and from 2 to 32 potential facilities, and have compared the results with general purpose global optimization solvers (BARON, Antigone and SCIP). The results show that the bilevel decomposition algorithm was able to find the optimal solution within 1% optimality tolerance in 87% of the case studies, and performed better than the other general purpose optimization solvers in 73% of them. We have noticed that the improvement in performance due to the use of the algorithm is more apparent for larger instances, specifically the networks with larger number of suppliers and customers fixed points. The global optimization solver that performed the best for this type of problem was BARON. Antigone was the global solver that had the worst performance, not being able to find a feasible solution in 47% of the test cases.

7.1.2 Global Optimization Algorithm for Multi-period Design and Planning of Centralized and Distributed Manufacturing Networks

Chapter 3 highlighted the need for a general model to optimize the design and planning of Distributed and/or Centralized manufacturing networks considering potential trade-offs between capital and transportation costs over a given time horizon. We have proposed a GDP formulation to solve this problem, which is a multi-period extension of the model proposed in chapter 2.

We show that with the added complexity of having multi-period decisions the original Bilevel Decomposition proposed in chapter 2 and the available global optimization solvers (BARON, ANTIGONE and SCIP) do not perform well, taking a long time to find feasible solutions and an acceptable optimality gap. Therefore, we have proposed an accelerated version of the Bilevel Decomposition with additional steps: Facility Pruning, Partition Pruning and Warm-start of the Master Problem. We show that the additional steps do not compromise the rigorousness of the algorithm, which still has ϵ-convergence as proved in chapter 2. Additionally, we discuss theoretical properties of the algorithm and find an upper bound to the space discretization such that if the space is partitioned in any finer grid, the algorithm is guaranteed to converge in a single iteration.

Additionally, we have performed computational experiments for the multi-period version of 10 random instances from chapter 2. The results show that the Accelerated Bilevel Decomposition algorithm was able to find the optimal solution within 2% optimality tolerance in 70% of the case studies, and performed better (i.e. found the optimal faster) than the other options in all of them. It should be noted that there was a considerable improvement in performance between the original Bilevel Decomposition and our Accelerated version of it, being able to solve 7 out of 10 instances instead of 5 out of 10. The global optimization solver that had the best performance for this problem and these instances was BARON. SCIP and ANTIGONE had a similar performance, only being able to solve 2 out of the 10 instances.

Finally, we have illustrated the applicability of the model and algorithm by solving a biomass supply chain problem from the literature (Lara and Grossmann, 2016; Chen and Grossmann, 2019). The resulting model has 3,457 constraints, 1,545 continuous variables, and 1,320 binary variables. Starting with $p_x, p_y = 5$ partitions and $N_x, N_y = 2$ increments it takes 3 iterations and 6 hours to solve it with the Accelerated Bilevel Decomposition within 2% optimality gap, which yields an optimal value of \$2.178 billion. We have attempted to solve this same instance with BARON, but it only achieved 68% optimality gap when it reached the maximum solution time of 10 hours, highlighting again the need for a specialized algorithm such as the proposed Accelerated Bilevel Decomposition to be able to solve real-world applications of this problem. The results show that all the 10 distributed facilities were built in year 1, and one centralized facility was built in year 2.

7.1.3 Impact of Model Resolution on Scenario Outcomes for Electric Power Generation Expansion

In chapter 4, we have performed in collaboration with researchers from ExxonMobil a systematic comparison of two alternate Generation Expansion Planning (GEP) frameworks to demonstrate how the choice of representing grid operations within a power system GEP framework can impact future projections of grid evolution as well as operational metrics such as unmet demand, curtailment, and renewable energy penetration. This preliminary analysis, which has not focused on algorithmic optimization, is instead focused on identifying trends based on a power system that approximately represents the U.S. Electric Reliability Council of Texas (ERCOT) grid in 2015, and motivates the more complex models proposed in the following chapters. The two least-cost power generation GEP models developed in this study, C-GEP and TS-GEP, are deterministic inter-temporal optimization models that take the viewpoint point of a centralized planner seeking to determine cost-optimal expansion decisions over a planning horizon of several decades. Temporally, the C-GEP represents annual load as well as wind and solar generation using up to 12 representative days at an hourly time resolution, whereas the TS-GEP represents annual load as well as wind and solar generation, with 16 time slices representing different times of day and seasons. It is important to note that the C-GEP model used in this chapter is simpler than the ones proposed in chapters 5 and 6 as it has a single load node (i.e. no transmission constraints), does not include the option of adding storage units, and does not handle uncertainty.

For the same set of technology and cost assumptions, we have found that the TS-GEP model is in general likely to overestimate solar PV capacity (by 35% in the case study presented) and underestimate wind and the supporting NG capacity requirements, compared the C-GEP model. This finding is explained primarily by the limited representation of the temporal variability in renewable energy generation, notably wind, and its correlation with load when using the time slice approach as compared to the chronological approach. For solar PV, using values of capacity factors based on 4-hr seasonal averages (as in the TS-GEP) overvalues the coincidence between peak solar PV generation and peak system load (also a 4-hr seasonal average) and consequently underestimates the declining value of solar PV generation with increasing penetration, as compared to the chronological approach using 12 representative days (as in the C-GEP) at an hourly resolution. The differences in the capacity mix to achieve the same renewable energy targets have reliability implications, as reflected by the lower unmet demand projected for the C-GEP capacity mix when tested in a detailed hourly simulation of annual grid operations.

While it is common for policy-focused GEP studies to test the capacity mix estimated by a GEP through a Production Cost Simulation (PCS) framework (Brinkman, 2015; Lew et al., 2013), our study has highlighted the importance of evaluating the operational performance of the capacity mix projections for multiple years of load and renewable energy generation profiles. Such analysis benchmarks the ability of the capacity mix to achieve the desired reliability and/or environmental attributes. For instance, the results presented here suggest that the unmet demand resulting from the capacity mix estimated by the C-GEP (using 12 representative days) is less sensitive to the annual variations in load and renewable energy generation profiles compared to the outputs projected by the TS-GEP. It is also worth noting that, although the unmet demand for the C-GEP capacity mix increases with increasing renewable energy targets, the maximum value is still relatively small at 0.023% of load for the 70% renewables scenario, which is comparable to the loss of load threshold values considered in estimating reserve margin requirements (Pfeifenberger et al., 2013).

Even within a C-GEP framework, selecting fewer than 4 sample days may lead to considerable overestimation of solar PV capacity. This finding has implications for the choice of temporal resolution in not just power sector planning models, but also more broadly for multi-sector, multi-country energy economic and integrated assessment models. For example, it was recently suggested that the current time slice implementation in the electricity grid planning implementation of the 2016 NEMS energy-economic model for the US may be overestimating solar PV capacity projections (Wood, 2016). Similarly, Bistline et al. (2017) performed an intra-model comparison of alternative temporal representations in the US-REGEN model and concluded that using a seasonal-average approach (akin to TS-GEP) is likely to overestimate renewables capacity and underestimate investment in dispatchable generation, compared to the representative hours approach (akin to C-GEP). Our study contributes to the growing body of evidence on the need for using a temporal representation based on a few representative days or other parameterizations that yield similar behavior in multi-sector, energy-economic models and other energy system models supporting policy analysis and decision-making.

7.1.4 Deterministic Electric Power Infrastructure Planning: Mixed-integer Programming Model and Nested Decomposition Algorithm

In chapter 5 we have proposed a comprehensive Generation Expansion Planning optimization modeling framework, more general than the one proposed in chapter 4 to optimize the changes in the power systems infrastructure required to meet the projected electricity demand over the next few decades, while taking into account detailed operating constraints and the variability and intermittency of renewable generation sources. The Mixed-integer Linear Programming (MILP) model takes the viewpoint of a central planning entity whose goal is to identify the source (nuclear, coal, natural gas, wind and solar), generation and storage technology (e.g., steam, combustion and wind turbines, photovoltaic and concentrated solar panels, and batteries), location (regions), and capacity of future power generation technologies that can meet the projected electricity demand, while minimizing the net present cost, which includes investment, operating, and environmental costs (e.g. carbon tax and renewable generation quota). The proposed model uses a reduced network, does not allow transmission expansion, and assumes that the flow in each transmission line can be determined by an energy balance between nodes, which ignores Kirchhoff's voltage law. The major contributions of this model are in the description of retirement, inclusion of storage, and handling of multi-scale aspects with clustering and time scale strategies.

In order to solve this large-scale multi-period MILP problem more efficienttly, we have developed a decomposition algorithm based on Nested Benders Decomposition for mixed-integer multi-period problems to solve large-scale models. This framework was originally developed for stochastic programming by Zou et al. (2018b), but we have adapted it to deterministic multi-period problems. We have modified it to handle integer and continuous state variables, at the expense of sacrificing the finite convergence property due to potential duality gap, although in the implementation a maximum number of iterations is specified. We have also proved the validity of the Benders, Strengthened Benders and Lagrangean cuts for handling mixed-integer recourse, and have applied acceleration techniques to improve the overall performance of the algorithm.

The formulation and solution framework were tested for a case study in the ERCOT region. We first solved the proposed model for 1 to 12 representative days to assess the impact of the number of representative days in the planning strategy. This is different than the analysis from chapter 4 because the model here has transmission constraints and 5 load nodes (note that this case study assumes no storage expansions). The results show that the wind capacity suffers minor fluctuations and slowly converges to a capacity that is 3% higher than the 1-day model projected. On the other hand, a major difference occurs in projected solar photovoltaic (PV) and natural gas (NG) generation capacities. For 1 to 4 representative days, there is a clear trend of decreasing PV and increasing NG contributions as we increase the number of representative days. With the exception of the outlier in the "5 representative days" results, this trend can be seen for all the variants until the results reach a plateau with small fluctuations, which is in agreement with the findings in chapter 4.

We have also compared the impact of the having more representative days in the cost breakdown. The results indicate that for both cases the cost is driven by fuel consumption (coal, NG and uranium), which accounts for about 60% of the total cost. The major difference between these representations is in the startup contribution, which goes from 0.3% to 1.5% of the total cost when considering 4 to 12 representative days. These results show that by having a better representation of the dynamics of the systems, i.e., higher number of representative days with hourly load and capacity factor, the startup cost becomes more relevant.

We have tested the Nested Decomposition algorithm performance for 4 and 12 representative days per year. The full-space MILP model using 4 representative days has 1,201,761 constraints, 413,644 discrete variables, and 594,147 continuous variables. We have solved the problem using several methods: (1) we solved the monolithic formulation directly in CPLEX; (2) we applied the Nested Decomposition algorithm with each of the cut options; (3) we employed the acceleration technique to the Nested Decomposition variant with the best performance in (2). The results show that all of the versions of the Nested Decomposition algorithm found solutions within 10% gap in less than 10 minutes, while the first integer solution by the monolithic MILP took 2.04 hours. The Benders cuts were the most efficient among the possible cuts, finding a solution within 1% gap in 47.5 minutes. Moreover, the accelerated Nested Decomposition with Benders cuts greatly reduced the optimality gap in the initial iterations and is the fastest (39.6 minutes) at finding a solution within 1% optimality gap.

The full-space MILP model for the 12 representative days variant has 3,626,721 constraints, 1,243,084 discrete variables, and 1,774,947 continuous variables. We have solved the problem using the Nested Decomposition algorithm with the two versions that performed the best using 4 representative days (Benders cuts, and warm-start cuts + Benders cuts). Both versions of the algorithm obtained solutions within 2% optimality gap in less than 5 hours. We notice again that by "warm-starting" the algorithm we were able to reduce the optimality gap in the initial iterations. The accelerated version found a solution with a 2% gap in its 4^{th} iteration and 3.5 hours, while the normal Benders cut version took 4.6 hours and 7 iterations to reach the same optimality gap. These improvements in solution time are important because they allow one to perform several sensitivity analysis to better understand the drivers for a variety of scenarios.

7.1.5 Electric Power Infrastructure Planning under Uncertainty: Stochastic Dual Dynamic Integer Programming (SDDiP) and parallelization scheme

In chapter 6 we have proposed a Multistage Stochastic Mixed-Integer programming (MSIP) formulation to address the long-term generation expansion planning under uncertainty, which is an extension of the deterministic MILP model proposed in chapter 5. In our framework we capture both long-scale strategic and short-scale operating (i.e., multiple hourly profiles of representative days capturing load demand and renewable energy capacity factor) uncertainties. The strategic uncertainties occur in the same time scale as the investment decisions (e.g. fuel price and carbon tax), while the operating uncertainties are captured through different representative days' profiles of hourly load demand and renewable availability between investment decisions.

We have decomposed the model using Stochastic Dual Dynamic integer Programming (SDDiP) algorithm, which exploits the nested structure in this MSIP problem. The SDDiP relies on the stage-wise independence assumption and the ability to share cuts among different nodes in the same stage to avoid the combinatorial explosion. Therefore, it provides a practical solution for solving real-world applications of MSIP on very large scenario trees without the need for scenario reduction methods. Additionally, we have proposed a parallelization scheme that takes advantage of the fact that in the SDDiP framework the subproblems of the nodes within the same stage are independent from each other.

We have tested the proposed MSIP formulation and SDDiP algorithm for a case study in the ERCOT region. This case study is based on the deterministic case study presented in chapter 5, with the addition of operational and strategic uncertainties, and the option of adding storage units. We first tested the parallel efficiency of our algorithm by solving the same 5-stage problem both sequentially and in parallel. The solution time for the serial implementation is $t_s = 13,950$ seconds and for the parallel implementation with 3 processes is $t_p = 6,131$. Therefore, our parallel SDDiP algorithm has 76% efficiency.

To test the capabilities of our algorithm, we have solved the problem for: (i) 5 stages (5 years, 1 year per stage); (ii) 10 stages (10 years, 1 year per stage); and (iii) 15 stages (15 years, 1 year per stage). In all of them, the subproblem (node) size before cuts is 50,042 constraints, 13,746 integer variables, and 22,755 continuous variables. The extensive forms of all three cases are massive, varying from the order of 10^7 variables and constraints (in the 5-stage problem) to 10^{15} variables and constraints (in the 15-stage problem). Considering the size of the models, all cases were solved in reasonable amounts of time: 1.7 hours, 2.3 hours and 23.9 hours, respectively. Due to memory limitations, it is fair to say that at least the 10-stage and 15-stage cases would not be solvable on a personal laptop or desktop without decomposing the model, or if the model is decomposed by a scenario-based approach without scenario reduction techniques. Additionally, the convergence time is considerably shorter than similar sized GEP problems with multistage stochastic programming formulations reported in the literature that use scenario-based decomposition (e.g. Liu et al. (2018)). These results confirm how powerful and useful SDDiP can be for practical large-scale MSIP models.

The results also show that for most of the scenario trees tested (with natural gas price and carbon tax uncertainty) the first-stage decisions consist of investing in new natural gas plants, indicating the competitiveness of this source for this case study. We also have run a case study in which all nuclear power generation is immediately retired, and unless we consider steep values for the high realization of carbon tax the optimization still decides to invest in natural gas turbines in the first stage. Finally, we show the value of stochastic solution for the scenario tree with high carbon tax uncertainty, with a potential $ 2 billion reduction in cost in the long run.

7.2 Research Contributions

The major contributions of this thesis can be summarized as follows:

1. Proposed an extension to the Capacitated Multi-facility Weber Problem that has fixed costs, multiple types of facilities, and two sets of fixed points representing suppliers and consumers. The problem was formulated as a nonconvex GDP, which was reformulated as a nonconvex MINLP.
2. Proposed a Bilevel Decomposition algorithm for the extended Capacitated Multi-facility Weber Problem that converges to the global optimum within an ϵ-tolerance in a finite number of iterations and outperfors commercial global solvers especially for larger instances.
3. Proposed a systematic framework to solve the optimal design and multi-period planning of centralized and distributed manufacturing networks, which capture the trade-offs between capital and transportation costs. The proposed model, which is an extension to the Capacitated Multi-facility Weber Problem, is a multi-period nonlinear GDP, reformulated as a multi-period nonconvex MINLP.

4. Proposed an accelerated version of the bilevel decomposition algorithm by Lara et al. (2018c) that keeps its rigor (i.e., its ϵ-convergence), but has some additional steps to improve its performance to allow the solution of large-scale multi-period instances of this problem within a reasonable amount of time. The additional steps are aimed at reducing the optimization search space such that it is easier for the Bilevel Decomposition to find good bounds and the optimal solution.
5. Compared the outputs of a Generation Expansion Planning (GEP) with chronological time-representation (C-GEP) of grid operations using 12 representative days against those of a traditional GEP, which uses a seasonal average or time-slice representation (TS-GEP) to mimic the salient features in models like IPM (EPA, 2015), ReEDS (Short et al., 2011), and the Electricity Market Module in NEMS (EIA, 2016). This inter-model comparison across multiple hypothetical scenarios of renewable energy penetration is an improvement over other recent multi-model comparisons (e.g. Cole et al. (2017); Bistline et al. (2017)).
6. Tested the robustness of the grid operations approximations made by both C-GEP and TS-GEP models by evaluating their projected capacity mix through a production cost simulation (PCS) model, which simulates annual grid operations at an hourly resolution. The proposed methodology illustrates the importance of evaluating operational outcomes associated with a GEP projected capacity mix when considering the prevailing variability in load and profiles, both within a year and across multiple years.
7. Proposed a comprehensive GEP optimization modeling framework to optimize the changes in the power systems infrastructure required (nuclear, coal, natural gas, wind and solar) to meet the projected electricity demand over the next few decades, while taking into account detailed operating constraints, the variability and intermittency of renewable generation sources, and potential inclusion of storage units. The major contributions of this model are in the description of retirement, inclusion of storage options, and handling of multi-scale aspects with clustering and time scale strategies.
8. Proposed a decomposition algorithm based on Nested Benders Decomposition for

mixed-integer multi-period problems to solve large-scale models. This framework, originally developed for stochastic programming by Zou et al. (2018b), was adapted to deterministic multi-period problems, and modified to handle integer and continuous state variables (at the expense of losing the finite convergence property due to potential duality gap), proved the validity of the Benders, Strengthened Benders and Lagrangean cuts for this case of mixed-integer recourse, and have applied acceleration techniques to improve the overall performance of the algorithm.

9. Tested the proposed deterministic GEP model and Nested Decomposition for a case study in the ERCOT region and investigated the impact of the number of representative days selected to represent a year of grid operations on resulting capacity and generation projections.
10. Extended the MILP model proposed by Lara et al. (2018a) to multistage stochastic mixed-integer programming with both strategic (yearly) and operational (hourly) uncertainty, for which we applied Stochastic Dual Dynamic Programming (SDDiP) with mixed-integer recourse in the context of GEP optimization with integrated operating decisions and proposed a parallelization scheme to solve the SDDiP algorithm more efficiently.
11. Applied of the stochastic GEP model and SDDiP algorithmic framework to a case-study in the ERCOT region for scenario trees considering natural gas price and carbon tax uncertainty for the reference case, and a hypothetical case without nuclear power. This study shows that the parallelized SDDiP algorithm allows the solution of multistage stochastic programming models with up to quadrillions of variables and constraints in reasonable amounts of time.

7.3 Papers produced from this dissertation

Lara, C.L., Trespalacios, F., Grossmann, I.E. *Global Optimization Algorithm for Capacitated Multi-facility Continuous Location Allocation Problems*, Journal of Global Optimization, 2018. 71:871–889.

Lara C.L., Bernal, D.E., Li, C., Grossmann I.E., *Global Optimization for Multi-period Design and Planning of Centralized and Distributed Manufacturing Networks*, Computers & Chemical Engineering, 2019. Submitted for publication.

Mallapragada, D., Papageorgiou, D., Venkatesh, A., Lara C.L., Grossmann I.E. *Impact of model resolution on scenario outcomes for electricity sector system expansion*, Energy, 2018. 163:1231-1244.

Lara C.L., Mallapragada, D., Papageorgiou, D., Venkatesh, A., Grossmann I.E., *Deterministic Electric Power Infrastructure Planning: Mixed-integer Linear Programming Model and Nested Decomposition*, European Journal of Operational Research, 2018. 271 (3):1037-1054.

Lara C.L., Siirola, J.D., Grossmann I.E., *Electric Power Infrastructure Planning Under Uncertainty: Stochastic Dual Dynamic Integer Programming (SDDiP) and parallelization scheme.*, Optimization and Engineering. 2019. Submitted for publication, 2019.

7.4 Future research directions

7.4.1 Continuous Location-Allocation Design and Planning

Allow relocation of the distributed modular facilities

An option that can be added to the multi-period design and planning of centralized and distributed facility networks proposed in chapter 3 is to allow the relocation of the distributed modular plants, subject to a transportation cost relative to the distance between the sites. Chen and Grossmann (2019) show that, in the context of their MILP model (which is similar to ours but assumes new facilities can only be located in a set of pre-defined potential locations), there is value in allowing module relocation especially in the presence of significant transportation costs and demand variability. Therefore, it would be interesting to see the effects of adding this flexibility in our 2-dimensional continuous location-allocation framework.

Allow forbidden areas inside the feasible 2-dimensional space

The current GDP formulations proposed in chapters 2 and 3 assume that the entire rectangle that includes the fixed points (suppliers and markets) is feasible for locating new manufacturing facilities. However, in reality, there are location limitations in the placement of new facilities (e.g. a lake, or a portion of land that is not for sale). Therefore, an option that could be added to both the single-period and the multi-period formulations is to have forbidden areas within the feasible 2-dimensional space. Depending on the shape of this forbidden areas (e.g. polyhedral, convex, nonconvex) this could greatly impact the computational difficulty of solving this problem, and adaptations to the Accelerated Bilevel Decomposition may be necessary.

Allow fixed and variable costs to be location dependent

The current formulation does not account for the changes in fixed and variable investment and operating costs due to the variable cost of real estate and utilities for different locations. This idea has been explored in the literature (Brimberg and Salhi, 2005) in the context of continuous location-allocation with no restriction on the capacity of the facilities. It would be interesting to see how this would affect the results of our capacitated continuous location-allocation framework.

Use Satisfiability (SAT) to solve the Facility Pruning and Partition Pruning steps in the Accelerated Bilevel Decomposition

The Satisfiability problem (SAT) is a constraint satisfaction technique and consists of determining if there exists an interpretation of a given problem that satisfies a given Boolean formula (Mistry et al., 2018). If the Boolean formula can be consistently satisfied, it evaluates to TRUE. However, if no such assignment exists, the function expressed by the formula is FALSE and the formula is unsatisfiable.

Instead of formulating the subproblems of the Facility Pruning and Partition Pruning steps as MILPs, we can be formulate them as SAT problems, such that their objective values, $\Phi^{LB,k'}$ and $\Phi^{LB,p'}$, have to be less or equal than the current upper bound UB^{iter}. The idea is to leverage the strength and robustness of modern SAT solvers, and potentially speed-up the solution of the Accelerated Bilevel Decomposition algorithm.

Extend the model to two-stage stochastic programming to handle demand uncertainty

The current formulation is deterministic, hence it does not handle the uncertainties present in these type of supply chain (e.g. demand uncertainty). The multi-period GDP formulation proposed in chapter 3 can be extended to two-stage stochastic programming such that the design decisions (i.e., which facilities to build and in which location) are first-stage *here-and-now* decisions, and the allocation decisions are second-stage *wait-and-see* decisions, which are taken after the values of the uncertain parameters (random variables) are revealed, allowing recourse action.

This extension would directly impact the size of the model and, consequently, its tractability. Therefore, it would most likely require changes in the Accelerated Bilevel Decomposition algorithm such as combining it with a Benders-like decomposition to take advantage of the "L-shape" structure of the two-stage stochastic programmming model.

7.4.2 Electric Power Infrastructure Planning

Improve the transmission representation in the model and include the option for transmission expansion

Both the deterministic MILP proposed in chapter 5 and the multistage stochastic programming formulation proposed in chapter 6 use "truck-route" representation for the transmission network, which ignores Kirchhoff's voltage law, and do not consider the option of transmission expansion. As mentioned in those chapters, the transmission infrastructure affects both location and type of generation investment. Therefore, not including transmission expansion and disregarding Kirchhoff's Voltage Law could distort the planning results (Munoz et al., 2013).

A future direction for this work is to include transmission expansion with a more detailed representation of transmission in both the deterministic and the stochastic formulations. The rigorous way of representing transmission between generation and load nodes in the system is through optimal power flow models (e.g., Frank et al. (2012a,b)). However, the constraints in this representation are nonlinear and nonconvex, which would greatly impact the tractability of the models and limit the size of solvable instances. However, there are in the literature MILP approximations for transmission expansion planning that have efficient computational behavior and better represent the physics of the systems (e.g. Alguacil et al. (2003)). Therefore, it would be interesting to include such constraints in the model and investigate how this addition would affect the planning strategy.

Perform sensitivity analysis of key parameters

It would be interesting to perform a sensitivity analysis of the key parameters in the GEP formulation such as planning reserve margin and annualized investment cost of new technologies (e.g. batteries, solar photo-voltaic panels, and concentrated solar panel) to see how much they influence the planning strategy.

Adapt the formulation to address the issue of deregulated markets

The proposed GEP framework is monopolistic and assumes a centralized planning entity. However, with the deregulation of the electricity markets, decisions are taken separately by the participating companies. It would be interesting to reformulate the GEP problem to represent competitive market behavior between generation entities/participants using game theory and multi-level optimization (e.g., Roh et al. (2007); Pozo et al. (2013); Cadre et al. (2015)).

Adapt the GEP formulation to include reliability

An important issue that is not included in the current formulation is the reliability of the system (i.e. to account for potential failure and outage of generators and transmission lines). As future work, the proposed GEP formulation can be adapted to include reliability issues, following what has been already proposed in the literature (Ballireddy and Modi, 2017).

Adapt the parallel SDDiP algorithm to handle stage-wise dependent parameters

The SDDiP relies on the stage-wise independence assumption and the ability to share cuts among different nodes in the same stage to avoid the combinatorial explosion. However, this assumption is not always valid depending on the uncertain parameters considered (e.g., learning rate of the new technologies). There has been some work on how to extend the use of SDDP/SDDiP for certain types of interstage dependency (Infanger and Morton, 1996; Shapiro et al., 2013a; de Queiroz and Morton, 2013; Lohmann et al., 2016) which could be used in our framework.

Additionally, there is the option of using the framework proposed by Rebennack (2016), which combines SDDP with the sampling-based stochastic nested Benders decomposition approach such that it can handle strategic uncertainty with stage-wise dependence and operating uncertainty with stage-wise independence.

Include construction lead time in the multistage stochastic programming formulation

In deterministic models it is common practice to assume *no lead time* between the decision to build/install a generator and the moment it can begin producing electricity, as one can subtract the construction time to the beginning-of-operations time given by the optimization model to find out when to initiate the power plants' construction. However, in a multistage stochastic framework this could have an impact in the planning strategy, especially if the lead time is included as an uncertain parameter. Therefore, it would be interesting to consider the lead time of construction of the power plants as both a deterministic and an uncertain parameter to evaluate how this would affect the planning strategy.

Adapt the parallel SDDiP algorithm to address risk-averse GEP problems

The classical Stochastic Programming formulation minimizes the expected cost and is risk neutral. However, while a risk-neutral approach may yield solutions that are good in the long run, it may perform poorly under certain realizations of the random data. For non-repetitive decision making problems under uncertainty, a risk-averse approach that considers the effects of the variability of random outcomes would provide more robust solutions compared to a risk-neutral approach (Noyan, 2012). Therefore introducing risk measures such as Conditional Value at Risk (CVaR) into the the proposed GEP is an interesting direction as the ultimate goal of such models is to plan a robust and reliable system that can supply enough electricity at all times.

There has been some effort in the literature to combine risk-averse approach with SDDP (Shapiro et al., 2013b). However, to the best of our knowledge, the combination of risk-averse approach with SDDiP has not yet been explored in the literature.

Evaluate and improve the parallel scalability of the SDDiP algorithm

In the implementation of the parallel SDDiP algorithm proposed in chapter 6, the nodes are statistically assigned to processes, and there is the potential for load imbalance due to both the random sample of scenarios and because of variance in the time to solve each MILP subproblem. Addressing this potential scaling issue is left for future work.

Bibliography

ABB. CREZ Reactive Power Compensation Study. Technical report, ABB Group, 2010.

S. Ahmed. Blessing of binary in stochastic integer programming. In *XIV International Conference on Stochastic Programming*, Jun. 25-30 2016.

M. H. Akyuz. Discretization based heuristics for the capacitated multi-facility weber problem with convex polyhedral barriers. *An International Journal of Optimization and Control: Theories & Applications (IJOCTA)*, 8(1):26–42, 2017. ISSN 2146-5703. doi: 10.11121/ijocta.01.2018.00388. URL `http://ijocta.balikesir.edu.tr/index.php/files/article/view/388`.

M. H. Akyuz, T. Oncan, and İ. K. Altinel. Solving the multi-commodity capacitated multi-facility Weber problem using lagrangean relaxation and a subgradient-like algorithm. *Journal of the Operational Research Society*, 63(6):771–789, 2012. ISSN 1476-9360. doi: 10.1057/jors.2011.81.

M. H. Akyuz, T. Oncan, and I. K. Altinel. Branch and bound algorithms for solving the multi-commodity capacitated multi-facility weber problem. *Annals of Operations Research*, Sep 2018. ISSN 1572-9338. doi: 10.1007/s10479-018-3026-5. URL `https://doi.org/10.1007/s10479-018-3026-5`.

I. M. Al-Loughani. *Algorithmic Approaches for Solving the Euclidean Distance Location and Location-Allocation Problems*. PhD thesis, Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, 1997.

M. Albadi and E. El-Saadany. Overview of wind power intermittency impacts on power systems. *Electric Power Systems Research*, 80(6):627 – 632, 2010. ISSN 0378-7796. doi: http://dx.doi.org/10.1016/j.epsr.2009.10.035. URL `http://www.sciencedirect.com/science/article/pii/S0378779609002764`.

V. M. Albornoz, P. Benario, and M. E. Rojas. A two-stage stochastic integer programming model for a thermal power system expansion. *International Transactions in Operational Research*, 11(3):243–257, 2004. doi: 10.1111/j.1475-3995.2004.00456.x.

N. Alguacil, A. L. Motto, and A. J. Conejo. Transmission expansion planning: a mixed-integer lp approach. *IEEE Transactions on Power Systems*, 18(3):1070–1077, Aug 2003. ISSN 0885-8950. doi: 10.1109/TPWRS.2003.814891.

D. Anderson, N. Samaan, T. Nguyen, and M. Kintner-Mayer. North America Modeling Compendium and Analysis. Technical report, Pacific Northwest National Laboratory, Richland, WA, 2016. URL `https://www.energy.gov/sites/prod/files/2017/01/f34/North%20America%20Modeling%20Compendium%20and%20Analysis.pdf`.

N. Aras, I. K. Altinel, and M. Orbay. New heuristic methods for the capacitated multi-facility Weber problem. *Naval Research Logistics (NRL)*, 54(1):21–32, 2007. ISSN 1520-6750. doi: 10.1002/nav.20176.

O. M. Babatunde, J. L. Munda, and Y. Hamam. Generation expansion planning: A survey. *2018 IEEE PES/IAS PowerAfrica*, pages 307–312, 2018.

L. Bahiense, G. C. Oliveira, M. Pereira, and S. Granville. A mixed integer disjunctive model for transmission network expansion. *IEEE Transactions on Power Systems*, 16(3):560–565, Aug 2001. ISSN 0885-8950. doi: 10.1109/59.932295.

G. A. Bakirtzis, P. N. Biskas, and V. Chatziathanasiou. Generation expansion planning by MILP considering mid-term scheduling decisions. *Electric Power Systems Research*, 86: 98 – 112, 2012. ISSN 0378-7796. doi: http://dx.doi.org/10.1016/j.epsr.2011.12.008. URL `http://www.sciencedirect.com/science/article/pii/S0378779611003178`.

E. Balas. Disjunctive programming. In P. Hammer, E. Johnson, and B. Korte, editors, *Discrete Optimization II*, volume 5 of *Annals of Discrete Mathematics*, pages 3 – 51. Elsevier, 1979. doi: https://doi.org/10.1016/S0167-5060(08)70342-X. URL `http://www.sciencedirect.com/science/article/pii/S016750600870342X`.

E. Balas. Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. *SIAM Journal on Algebraic Discrete Methods*, 6(3):466–486, 1985.

T. R. K. R. Ballireddy and P. K. Modi. Generation expansion planning considering reliability of the system: A review on various optimization techniques. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 2, pages 137–142, April 2017. doi: 10.1109/ICECA.2017.8212780.

L. Baringo and A. Baringo. A stochastic adaptive robust optimization approach for the generation and transmission expansion planning. *IEEE Transactions on Power Systems*, 33(1):792–802, Jan 2018. ISSN 0885-8950. doi: 10.1109/TPWRS.2017.2713486.

E. M. L. Beale. On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):173–184, 1955. doi: 10.1111/j.2517-6161.1955.tb00191.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1955.tb00191.x`.

A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376, Mar 2004. ISSN 1436-4646. doi: 10.1007/s10107-003-0454-y.

J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4(1):238–252, Dec 1962. ISSN 0945-3245. doi: 10.1007/BF01386316. URL `https://doi.org/10.1007/BF01386316`.

M. L. Bergamini, P. Aguirre, and I. Grossmann. Logic-based outer approximation for globally optimal synthesis of process networks. *Computers & Chemical Engineering*, 29(9):1914 –

1933, 2005. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2005.04.003. URL `http://www.sciencedirect.com/science/article/pii/S0098135405001006`.

D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004. doi: 10.1287/opre.1030.0065.

T. Bieringer, S. Buchholz, and N. Kockmann. Future production concepts in the chemical industry: Modular - small-scale - continuous. *Chemical Engineering & Technology*, 36(6): 900–910, 2013. doi: 10.1002/ceat.201200631. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ceat.201200631`.

J. R. Birge. Decomposition and partitioning methods for multistage stochastic linear programs. *Operations Research*, 33(5):989–1007, 1985. doi: 10.1287/opre.33.5.989. URL `http://dx.doi.org/10.1287/opre.33.5.989`.

J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Publishing Company, Incorporated, 2nd edition, 2011. ISBN 1461402360, 9781461402367.

J. Bistline, D. Shawhan, G. Blanford, F. de la Chesnaye, B. Mao, N. Santen, R. Zimmerman, and A. Krupnick. Systems Analysis in Electric Power Sector Modeling: Evaluating Model Complexity for Long-Range Planning. Technical report, Electric Power Research Institute (EPRI) and Resources For the Future, 2017.

Black & Veatch. Cost and performance data for power generation technologies. Technical report, Black & Veatch Holding Company, 2012. URL `https://www.bv.com/docs/reports-studies/nrel-cost-report.pdf`.

G. J. Blanford, J. H. Merrick, J. E. T. Bistline, and D. T. Young. Simulating annual variation in load, wind, and solar by representative hour selection. *Energy Journal*, 39(3): 189 – 212, 2018. ISSN 01956574. URL `https://ideas.repec.org/a/aen/journl/ej39-3-blanfor.html`.

J. Brimberg and S. Salhi. A continuous location-allocation problem with zone-dependent fixed cost. *Annals of Operations Research*, 136(1):99–115, 2005. ISSN 1572-9338. doi: 10.1007/s10479-005-2041-5.

J. Brimberg, P. Hansen, N. Mladonovic, and S. Salhi. A survey of solution methods for the continuous location allocation problem. *International Journal of Operational Research*, 5 (1):1–12, 2008.

G. Brinkman. Renewable Electricity Futures: Operational Analysis of the Western Interconnection at Very High Renewable Penetrations. Technical Report September, National Renewable Technology Laboratory (NREL), Golden, CO, 2015. URL `https://www.nrel.gov/docs/fy15osti/64467.pdf`.

S. Bruno, S. Ahmed, A. Shapiro, and A. Street. Risk neutral and risk averse approaches to multistage renewable investment planning under uncertainty. *European Journal of Operational Research*, 250(3):979 – 989, 2016. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2015.10.013.

H. L. Cadre, A. Papavasiliou, and Y. Smeers. Wind farm portfolio optimization under network capacity constraints. *European Journal of Operational Research*, 247(2):560 – 574, 2015. ISSN 0377-2217. doi: http://dx.doi.org/10.1016/j.ejor.2015.05.080. URL `http://www.sciencedirect.com/science/article/pii/S0377221715004920`.

B. A. Calfa. *Data Analytics Methods for Enterprise-wide Optimization under Uncertainty*. PhD thesis, Carnegie Mellon University, 2015.

C. C. Carøe and R. Schultz. Dual decomposition in stochastic integer programming. *Operations Research Letters*, 24(1):37 – 45, 1999. ISSN 0167-6377. doi: https://doi.org/10.1016/S0167-6377(98)00050-9.

S. Ceria and J. Soares. Convex programming for disjunctive convex optimization. *Mathematical Programming*, 86(3):595–614, 1999.

S. Cerisola, A. Baíllo, J. M. Fernández-López, A. Ramos, and R. Gollmer. Stochastic power generation unit commitment in electricity markets: A novel formulation and a comparison of solution methods. *Operations Research*, 57(1):32–46, 2009. doi: 10.1287/opre.1080.0593. URL `http://dx.doi.org/10.1287/opre.1080.0593`.

R. Chandra, L. Dagum, D. Kohr, D. Maydan, J. McDonald, and R. Menon. *Parallel Programming in OpenMP*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001. ISBN 1-55860-671-8, 9781558606715.

C. Chen, Y. Li, G. Huang, and Y. Li. A robust optimization method for planning regional-scale electric power systems and managing carbon dioxide. *International Journal of Electrical Power & Energy Systems*, 40(1):70 – 84, 2012. ISSN 0142-0615. doi: https://doi.org/10.1016/j.ijepes.2012.02.007.

J. S. Chen, S. Pan, and C. H. Ko. A continuation approach for the capacitated multi-facility Weber problem based on nonlinear SOCP reformulation. *Journal of Global Optimization*, 50(4):713–728, 2011. ISSN 09255001. doi: 10.1007/s10898-010-9632-7.

Q. Chen and I. E. Grossmann. Effective GDP optimization models for modular process synthesis. *Submitted for publication*, 2019.

W. Cole, B. Frew, T. Mai, Y. Sun, J. Bistline, G. Blanford, D. Young, C. Marcy, C. Namovicz, R. Edelman, and B. Meroney. Variable Renewable Energy in Long-Term Planning Models: A Multi-Model Perspective. Technical Report November, National Renewable Technology Laboratory, Electric Power Research Institute, U.S. Energy Information Administration, U.S. Environmental Protection Agency and U.S. Department of Energy, Golden, CO, 2017. URL `https://www.nrel.gov/docs/fy18osti/70528.pdf`.

A. J. Conejo, E. Castillo, R. Mínguez, and R. García-Bertrand. *Decomposition Techniques in Mathematical Programming: Engineering and Science Applications*. Springer Science+Business Media., 2006. New York, NY. USA.

M. Conforti, G. Cornuéjols, and G. Zambelli. *Reformulations and Relaxations*, pages 321–350. Springer International Publishing, Cham, 2014. ISBN 978-3-319-11008-0. doi: 10.1007/978-3-319-11008-0_8. URL https://doi.org/10.1007/978-3-319-11008-0_8.

L. Cooper. The transportation-location problem. *Operations Research*, 20(1):94–108, 1972.

L. Cooper. The fixed charge problem-i: A new heuristic method. *Computers & Mathematics With Applications - COMPUT MATH APPL*, 1:89–95, 01 1975. doi: 10.1016/0898-1221(75)90010-3.

L. Cooper. An efficient heuristic algorithm for the transportation-location problem. *Journal of Regional Science*, 16(3):309, 1976. ISSN 00224146.

G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3/4):197–206, 1955. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2627159.

A. R. de Queiroz and D. P. Morton. Sharing cuts under aggregated forecasts when decomposing multi-stage stochastic programs. *Operations Research Letters*, 41(3):311 – 316, 2013. ISSN 0167-6377. doi: https://doi.org/10.1016/j.orl.2013.03.003.

F. J. de Sisternes and M. D. Webster. Optimal selection of sample weeks for approximating the net load in generation planning problems. Technical report, MIT Energy Systems Division, 2013. URL https://dspace.mit.edu/handle/1721.1/102959.

F. J. de Sisternes, J. D. Jenkins, and A. Botterud. The value of energy storage in decarbonizing the electricity sector. *Applied Energy*, 175:368 – 379, 2016. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2016.05.014. URL http://www.sciencedirect.com/science/article/pii/S0306261916305967.

J. Deane, A. Chiodi, M. Gargiulo, and B. P. O. Gallachóir. Soft-linking of a power systems model to an energy systems model. *Energy*, 42(1):303 – 312, 2012. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2012.03.052. URL http://www.sciencedirect.

com/science/article/pii/S0360544212002551. 8th World Energy System Conference, WESC 2010.

T. A. Deetjen, J. B. Garrison, J. D. Rhodes, and M. E. Webber. Solar pv integration cost variation due to array orientation and geographic location in the electric reliability council of texas. *Applied Energy*, 180:607 – 616, 2016. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2016.08.012. URL http://www.sciencedirect.com/science/article/pii/S0306261916310984.

D. Dentcheva and W. Römisch. Optimal power generation under uncertainty via stochastic programming. In K. Marti and P. Kall, editors, *Stochastic Programming Methods and Technical Applications*, pages 22–56, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-642-45767-8.

A. Diamant. The Electric Generation Expansion Analysis System (EGEAS) Software. Technical report, Electric Power Research Institute, 10 2017.

J. Ding and A. Somani. A long-term investment planning model for mixed energy infrastructure integrated with renewable energy. In *2010 IEEE Green Technologies Conference*, pages 1–10, April 2010. doi: 10.1109/GREEN.2010.5453785.

DOE. Business Energy Investment Tax Credit (ITC). http://energy.gov/savings/business-energy-investment-tax-credit-it, 2016a. Online; Accessed 16 October 2016.

DOE. Renewable Electricity Production Tax Credit (PTC). http://energy.gov/savings/renewable-electricity-production-tax-credit-ptc, 2016b. Online; Accessed 11 October 2016.

E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002. ISSN 1436-4646. doi: 10.1007/s101070100263.

EIA. Annual Energy Outlook 2015: with projections to 2040. Technical report, U.S. Energy Information Administration (EIA), 2015. URL https://www.eia.gov/outlooks/aeo/pdf/0383(2015).pdf.

EIA. Annual Energy Outlook 2016: with projections to 2040. Technical report, U.S. Energy Information Administration, Washington, DC, 2016. URL https://www.eia.gov/outlooks/aeo/pdf/0383(2016).pdf.

EIA. Annual Energy Outlook 2019. Technical report, U.S. Energy Information Administration, 2019. URL https://www.eia.gov/outlooks/aeo/pdf/aeo2019.pdf.

EPA. Documentation for EPA Base Case v5.13 Using the Integrated Planning Model. Technical report, U.S. Environmental Protection Agency, Washington, DC, 2013. URL https://www.epa.gov/sites/production/files/2015-07/documents/documentation_for_epa_base_case_v.5.13_using_the_integrated_planning_model.pdf.

EPA. Incremental Documentation for EPA Base Case v.5.15 using the Integrated Planning Model. Technical report, U.S. Environmental Protection Agency, Washington, DC, 2015. URL https://www.epa.gov/sites/production/files/2015-08/documents/epa_base_case_v.5.15_incremental_documentation_august_2015.pdf.

EPRI. PRISM 2.0: Regional Energy and Economic Model Development and Initial Application; US- REGEN Model Documentation. Technical report, Electric Power Research Institute, 2013. URL https://www.epri.com/#/pages/product/000000003002000128/.

EPRI. US-REGEN Unit Commitment Model Documentation. Technical report, Electric Power Research Institute (EPRI), Palo Alto, CA, 2015. URL https://www.epri.com/#/pages/product/3002004748/?lang=en-US.

ERCOT. Report on the Capacity, Demand and Reserves (CDR) in the ERCOT region, 2016-2025. Technical report, Electric Reliability Council of Texas, 2015a. URL http://www.ercot.com/content/gridinfo/resource/2015/adequacy/cdr/CapacityDemandandReserveReport-December2015.pdf.

ERCOT. 2016 ERCOT System Planning: Long-Term Hourly Peak Demand and Energy Forecast. Technical report, Electric Reliability Council of Texas, 2015b. URL http://www.ercot.com/content/wcm/lists/143010/2018_Long-Term_Hourly_Peak_Demand_and_Energy_Forecast_Final.pdf.

ERCOT. Report on the Capacity, Demand, and Reserves (CDR) in the ERCOT Region, 2017-2026. Technical report, Electric Reliability Council of Texas (ERCOT), 2016. URL http://www.ercot.com/content/wcm/lists/96607/CapacityDemandandReserveReport-Dec2016.pdf.

ERCOT. Hourly Load Data Archives. http://www.ercot.com/gridinfo/load/load_hist/, 2016a. Online; 10 October 2016.

ERCOT. Resource Adequacy 2015 -CDR Peak Ave Wind Capacity Percentages, 2014-15 Winter Update. http://www.ercot.com/content/gridinfo/resource/2015/windsolar/CDR_PeakAveWindCapacityPercentages_04-27-2015.xls, 2016b. Online; Accessed 6 September 2016.

ERCOT. Long Term Study Task Force Meeting - Generic Database Characteristics REV 1. http://www.ercot.com/calendar/2011/5/3/34168-LTS, 2016c. Online; Accessed 07 September 2016.

ERCOT. Weather, 2016. URL http://www.ercot.com/about/weather.

ERCOT. Resource Adequacy: Wind and Solar. http://www.ercot.com/gridinfo/resource, 2017. Online; Accessed 19 April 2017.

ERCOT Systems Planning. Panhandle Renewable Energy Zone (PREZ) Study Report. Technical Report April, Electric Reliability Council of Texas, Inc., 2014. URL http://www.ercot.com/content/news/presentations/2014/Panhandle%20Renewable%20Energy%20Zone%20Study%20Report.pdf.

M. Fischetti and A. Lodi. *Heuristics in Mixed Integer Programming*. American Cancer Society, 2011. ISBN 9780470400531. doi: 10.1002/9780470400531.eorms0376. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470400531.eorms0376.

M. L. Fisher. The lagrangian relaxation method for solving integer programming problems. *Manage. Sci.*, 50(12 Supplement):1861–1871, Dec. 2004. ISSN 0025-1909. doi: 10.1287/mnsc.1040.0263. URL http://dx.doi.org/10.1287/mnsc.1040.0263.

A. Flores-Quiroz, R. Palma-Behnke, G. Zakeri, and R. Moreno. A column generation approach for solving generation expansion planning problems with high renewable energy penetration. *Electric Power Systems Research*, 136:232 – 241, 2016. ISSN 0378-7796. doi: http://dx.doi.org/10.1016/j.epsr.2016.02.011. URL http://www.sciencedirect.com/science/article/pii/S0378779616300177.

A. Frangioni. "about lagrangian methods in integer optimization". *Annals of Operations Research*, 139(1):163–193, 2005. ISSN 1572-9338. doi: 10.1007/s10479-005-3447-9. URL http://dx.doi.org/10.1007/s10479-005-3447-9.

S. Frank, I. Steponavice, and S. Rebennack. Optimal power flow: a bibliographic survey ii. *Energy Systems*, 3(3):259–289, 2012a. ISSN 1868-3975. doi: 10.1007/s12667-012-0057-x. URL http://dx.doi.org/10.1007/s12667-012-0057-x.

S. Frank, I. Steponavice, and S. Rebennack. Optimal power flow: a bibliographic survey i. *Energy Systems*, 3(3):221–258, 2012b. ISSN 1868-3975. doi: 10.1007/s12667-012-0056-y. URL http://dx.doi.org/10.1007/s12667-012-0056-y.

L. Gacitua, P. Gallegos, R. Henriquez-Auba, A. Lorca, M. Negrete-Pincetic, D. Olivares, A. Valenzuela, and G. Wenzel. A comprehensive review on expansion planning: Models and tools for energy policy analysis. *Renewable and Sustainable Energy Reviews*, 98: 346 – 360, 2018. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2018.08.043. URL http://www.sciencedirect.com/science/article/pii/S1364032118306269.

G. Gamrath, T. Fischer, T. Gally, A. M. Gleixner, G. Hendel, T. Koch, S. J. Maher, M. Miltenberger, B. Müller, M. E. Pfetsch, C. Puchert, D. Rehfeldt, S. Schenker, R. Schwarz, F. Serrano, Y. Shinano, S. Vigerske, D. Weninger, M. Winkler, J. T. Witt, and J. Witzig. The SCIP Optimization Suite 3.2. Technical Report 15-60, ZIB, Takustr.7, 14195 Berlin, 2016.

W. Gandulfo, E. Gil, and I. Aravena. Generation capacity expansion planning under demand uncertainty using stochastic mixed-integer programming. In *2014 IEEE PES General Meeting | Conference Exposition*, pages 1–5, July 2014. doi: 10.1109/PESGM.2014.6939368.

GE Power. Gas Power Systems Catalog. Technical report, General Electric Company, 2016.

A. Geoffrion. Lagrangian relaxation and its uses in integer programming. *Mathematical Programming*, 2, 01 1974.

E. Gil, I. Aravena, and R. Cárdenas. Generation capacity expansion planning under hydro uncertainty using stochastic mixed integer programming and scenario reduction. *IEEE Transactions on Power Systems*, 30(4):1838–1847, July 2015. ISSN 0885-8950. doi: 10.1109/TPWRS.2014.2351374.

A. Gleixner, L. Eifler, T. Gally, G. Gamrath, P. Gemander, R. L. Gottwald, G. Hendel, C. Hojny, T. Koch, M. Miltenberger, B. Müller, M. E. Pfetsch, C. Puchert, D. Rehfeldt, F. Schlösser, F. Serrano, Y. Shinano, J. M. Viernickel, S. Vigerske, D. Weninger, J. T. Witt, and J. Witzig. The SCIP Optimization Suite 5.0. Technical report, Optimization Online, December 2017. URL `http://www.optimization-online.org/DB_HTML/2017/12/6385.html`.

B. G. Gorenstin, N. M. Campodonico, J. P. Costa, and M. V. F. Pereira. Power system expansion planning under uncertainty. *IEEE Transactions on Power Systems*, 8(1):129–136, Feb 1993. ISSN 0885-8950. doi: 10.1109/59.221258.

C. E. Gounaris, R. Misener, and C. A. Floudas. Computational comparison of piecewise-linear relaxations for pooling problems. *Industrial & Engineering Chemistry Research*, 48 (12):5742–5766, 2009. doi: 10.1021/ie8016048.

I. E. Grossmann and S. Lee. Generalized convex disjunctive programming: Nonlinear convex hull relaxation. *Computational Optimization and Applications*, 26(1):83–100, 2003.

I. E. Grossmann and F. Trespalacios. Systematic modeling of discrete-continuous optimization models through generalized disjunctive programming. *AIChE Journal*, 59(9): 3276–3295, 2013.

I. E. Grossmann, R. M. Apap, B. A. Calfa, P. García-Herreros, and Q. Zhang. Recent advances in mathematical programming techniques for the optimization of process systems under uncertainty. *Computers & Chemical Engineering*, 91:3 – 14, 2016. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2016.03.002. 12th International Symposium on Process Systems Engineering & 25th European Symposium of Computer Aided Process Engineering (PSE-2015/ESCAPE-25), 31 May - 4 June 2015, Copenhagen, Denmark.

M. Guignard and S. Kim. Lagrangean decomposition: A model yielding stronger lagrangean bounds. *Mathematical Programming*, 39(2):215–228, Jun 1987. ISSN 1436-4646. doi: 10.1007/BF02592954. URL `https://doi.org/10.1007/BF02592954`.

V. Gupta and I. E. Grossmann. A new decomposition algorithm for multistage stochastic programs with endogenous uncertainties. *Computers & Chemical Engineering*, 62:62 – 79, 2014. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2013.11.011.

Gurobi Optimization. Gurobi optimizer reference manual, 2018. URL `http://www.gurobi.com`.

W. E. Hart, C. D. Laird, J.-P. Watson, D. L. Woodruff, G. A. Hackebeil, B. L. Nicholson, and J. D. Siirola. *Pyomo — Optimization Modeling in Python*, volume 67 of *Springer Optimization and Its Applications*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58819-3. doi: 10.1007/978-3-319-58821-6.

A. Helseth and H. Braaten. Efficient parallelization of the stochastic dual dynamic programming algorithm applied to hydropower scheduling. *Energies*, 8(12):14287–14297, 2015. ISSN 1996-1073. doi: 10.3390/en81212431.

C. F. Heuberger, I. Staffell, N. Shah, and N. M. Dowell. A systems approach to quantifying the value of power generation and energy storage technologies in future electricity networks. *Computers & Chemical Engineering*, 107:247 – 256, 2017. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2017.05.012. URL http://www.sciencedirect.com/science/article/pii/S0098135417302119. In honor of Professor Rafiqul Gani.

T. Homem-de Mello, V. L. de Matos, and E. C. Finardi. Sampling strategies and stopping criteria for stochastic dual dynamic programming: a case study in long-term hydrothermal scheduling. *Energy Systems*, 2(1):1–31, Mar 2011. ISSN 1868-3975. doi: 10.1007/s12667-011-0024-y.

K. Huang and S. Ahmed. The value of multistage stochastic programming in capacity planning under uncertainty. *Operations Research*, 57(4):893–904, 2009. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/25614804.

IBM. IBM ILOG CPLEX Optimization Studio CPLEX User's Manual. Technical report, IBM Corp., 2015. URL http://pic.dhe.ibm.com/infocenter/cosinfoc/v12r6/index.jsp?topic=/ilog.odms.ide.help/OPL_Studio/opllangref/topics/opl_langref_scheduling_sequence.html.

IEEFA. India's Electricity Sector Transformation. Technical report, Institute for Energy Economics and Financial Analysis, 2017. URL http://ieefa.org/wp-content/uploads/2017/11/India-Electricity-Sector-Transformation_Nov-2017-3.pdf.

G. Infanger and D. P. Morton. Cut sharing for multistage stochastic linear programs with interstage dependency. *Mathematical Programming*, 75(2):241–256, Nov 1996. ISSN 1436-4646. doi: 10.1007/BF02592154.

C. A. Irawan, M. Luis, S. Salhi, and A. Imran. The incorporation of fixed cost and multilevel capacities into the discrete and continuous single source capacitated facility location problem. *Annals of Operations Research*, Aug 2018. ISSN 1572-9338. doi: 10.1007/s10479-018-3014-9. URL `https://doi.org/10.1007/s10479-018-3014-9`.

S. Jin, S. M. Ryan, J.-P. Watson, and D. L. Woodruff. Modeling and solving a large-scale generation expansion planning problem under uncertainty. *Energy Systems*, 2(3):209–242, Nov 2011. ISSN 1868-3975. doi: 10.1007/s12667-011-0042-9.

S. Jin, A. Botterud, and S. M. Ryan. Temporal versus stochastic granularity in thermal generation capacity planning with wind power. *IEEE Transactions on Power Systems*, 29 (5):2033–2041, Sep. 2014. ISSN 0885-8950. doi: 10.1109/TPWRS.2014.2299760.

M. Kaut, K. T. Midthun, A. S. Werner, A. Tomasgard, L. Hellemo, and M. Fodstad. Multi-horizon stochastic programming. *Computational Management Science*, 11(1):179–193, Jan 2014. ISSN 1619-6988. doi: 10.1007/s10287-013-0182-6.

P. Y. Kerl, W. Zhang, J. B. Moreno-Cruz, A. Nenes, M. J. Realff, A. G. Russell, J. Sokol, and V. M. Thomas. New approach for optimal electricity planning and dispatching with hourly time-scale air quality and health considerations. *Proceedings of the National Academy of Sciences*, 112(35):10884–10889, 2015. doi: 10.1073/pnas.1413143112. URL `http://www.pnas.org/content/112/35/10884.abstract`.

H. W. Khun and R. E. Kuenne. An efficient algorithm for the commercial solution of the generallized Weber problem in spatial economics. *Journal of Regional Science*, 4(2):21–33, 1962.

K. Kim and V. M. Zavala. Algorithmic innovations and software for the dual decomposition method applied to stochastic mixed-integer programs. *Mathematical Programming Computation*, 10(2):225–266, Jun 2018. ISSN 1867-2957. doi: 10.1007/s12532-017-0128-z.

S. Kolodziej, P. M. Castro, and I. E. Grossmann. Global optimization of bilinear programs with a multiparametric disaggregation technique. *Journal of Global Optimization*, 57(4):

1039–1063, Dec 2013. ISSN 1573-2916. doi: 10.1007/s10898-012-0022-1. URL https://doi.org/10.1007/s10898-012-0022-1.

N. E. Koltsaklis and A. S. Dagoumas. State-of-the-art generation expansion planning: A review. *Applied Energy*, 230:563 – 589, 2018. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2018.08.087.

N. E. Koltsaklis and M. C. Georgiadis. A multi-period, multi-regional generation expansion planning model incorporating unit commitment constraints. *Applied Energy*, 158:310 – 331, 2015. ISSN 0306-2619. doi: http://dx.doi.org/10.1016/j.apenergy.2015.08.054. URL http://www.sciencedirect.com/science/article/pii/S0306261915009873.

V. Krishnan, J. Ho, B. F. Hobbs, A. L. Liu, J. D. McCalley, M. Shahidehpour, and Q. P. Zheng. Co-optimization of electricity transmission and generation resources for planning and policy analysis: review of concepts and modeling approaches. *Energy Systems*, 7(2): 297–332, 2016. ISSN 1868-3975. doi: 10.1007/s12667-015-0158-4.

E. Lannoye, D. Flynn, and M. O'Malley. The role of power system flexibility in generation planning. In *2011 IEEE Power and Energy Society General Meeting*, pages 1–6, July 2011. doi: 10.1109/PES.2011.6039009.

C. L. Lara. SDDiP implementation for a generation expansion planning model. https://github.com/cristianallara/SDDiP, 2019.

C. L. Lara and I. E. Grossmann. Global optimization for a continuous location-allocation model for centralized and distributed manufacturing. In *26th European Symposium on Computed Aided Process Engineering (ESCAPE)*, Portoroz, Slovenia, June 2016.

C. L. Lara, D. S. Mallapragada, D. J. Papageorgiou, A. Venkatesh, and I. E. Grossmann. Deterministic electric power infrastructure planning: Mixed-integer programming model and nested decomposition algorithm. *European Journal of Operational Research*, 271(3): 1037 – 1054, 2018a. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2018.05.039.

C. L. Lara, B. Omell, D. Miller, and I. E. Grossmann. Expanding the scope of electric power infrastructure planning. In M. R. Eden, M. G. Ierapetritou, and G. P. Towler, editors, *13th International Symposium on Process Systems Engineering (PSE 2018)*, volume 44 of *Computer Aided Chemical Engineering*, pages 1309 – 1314. Elsevier, 2018b. doi: https://doi.org/10.1016/B978-0-444-64241-7.50213-5.

C. L. Lara, F. Trespalacios, and I. E. Grossmann. Global optimization algorithm for capacitated multi-facility continuous location-allocation problems. *Journal of Global Optimization*, Feb 2018c. ISSN 1573-2916. doi: 10.1007/s10898-018-0621-6. URL https://doi.org/10.1007/s10898-018-0621-6.

C. Lassner. pymp. https://github.com/classner/pymp, 2018.

G. Latorre, R. D. Cruz, J. M. Areiza, and A. Villegas. Classification of publications and models on transmission expansion planning. *IEEE Transactions on Power Systems*, 18 (2):938–946, May 2003. ISSN 0885-8950. doi: 10.1109/TPWRS.2003.811168.

S. Lee and I. E. Grossmann. New algorithms for nonlinear generalized disjunctive programming. *Computers & Chemical Engineering*, 24(9):2125–2141, 2000.

T. Levin and A. Botterud. Electricity market design for generator revenue sufficiency with increased variable generation. *Energy Policy*, 87:392 – 406, 2015. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2015.09.012. URL http://www.sciencedirect.com/science/article/pii/S0301421515300999.

D. Lew, G. Brinkman, E. Ibanez, M. H. A. Florita, B.-M. Hodge, M. Hummon, G. Stark, J. King, S. Lefton, N. Kumar, D. Agan, G. Jordan, and S. Venkataraman. The Western Wind and Solar Integration Study Phase 2. Technical Report September, National Renewable Technology Laboratory (NREL), Golden, CO, 2013. URL https://www.nrel.gov/docs/fy13osti/55588.pdf.

S. Li. *Robust optimization of electric power generation expansion planning considering uncertainty of climate change*. PhD thesis, Rutgers, The State University of New Jersey, 2014.

S. Li, D. Coit, S. Selcuklu, and F. Felder. Electric power generation expansion planning: Robust optimization considering climate change. *IIE Annual Conference and Expo 2014*, pages 1049–1058, 01 2014.

Y. Liu, R. Sioshansi, and A. Conejo. Multistage stochastic investment planning with multiscale representation of uncertainties and decisions. *IEEE Transactions on Power Systems*, PP(99):1–1, 2017. ISSN 0885-8950. doi: 10.1109/TPWRS.2017.2694612.

Y. Liu, R. Sioshansi, and A. J. Conejo. Multistage stochastic investment planning with multiscale representation of uncertainties and decisions. *IEEE Transactions on Power Systems*, 33(1):781–791, Jan 2018. ISSN 0885-8950. doi: 10.1109/TPWRS.2017.2694612.

T. Lohmann, A. S. Hering, and S. Rebennack. Spatio-temporal hydro forecasting of multireservoir inflows for hydro-thermal scheduling. *European Journal of Operational Research*, 255(1):243 – 258, 2016. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2016.05.011.

J. A. Lopez, K. Ponnambalam, and V. H. Quintana. Generation and transmission expansion under risk using stochastic programming. *IEEE Transactions on Power Systems*, 22(3):1369–1378, Aug 2007. ISSN 0885-8950. doi: 10.1109/TPWRS.2007.901741.

R. Loulou, G. Goldstein, and K. Noble. Documentation for the MARKAL Family of Models. Technical report, International Energy Agency - Energy Technology Systems Analysis, 01 2004.

R. Loulou, U. Remne, A. Kanudia, A. Lehtila, and G. Goldstein. Documentation for the TIMES Model - PART I. Technical report, International Energy Agency - Energy Technology Systems Analysis, 01 2005.

M. Luis, S. Salhi, and G. Nagy. A constructive method and a guided hybrid GRASP for the capacitated multi-source Weber problem in the presence of fixed cost. *Journal of Algorithms and Computational Technology*, 9(2):215–232, 2015. doi: 10.1260/1748-3018.9.2.215.

M. Luis, M. F. Ramli, and A. Lin. A greedy heuristic algorithm for solving the capacitated planar multi-facility location-allocation problem. *AIP Conference Proceedings*, 1782(1): 040010, 2016. doi: 10.1063/1.4966077. URL `https://aip.scitation.org/doi/abs/10.1063/1.4966077`.

X. Luo, J. Wang, M. Dooner, and J. Clarke. Overview of current development in electrical energy storage technologies and the application potential in power system operation. *Applied Energy*, 137:511 – 536, 2015. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2014.09.081.

A. E. MacDonald, C. T. M. Clack, A. Alexander, A. Dunbar, J. Wilczak, and Y. Xie. Future cost-competitive electricity systems and their impact on US CO2 emissions. *Nature Climate Change*, 6:526–531, 2016. URL `https://doi.org/10.1038/nclimate2921`.

T. Mai, E. Drury, K. Eurek, N. Bodington, A. Lopez, and A. Perry. Resource Planning Model: An Integrated Resource Planning and Dispatch Tool for Regional Electric Systems. Technical Report January, National Renewable Energy Laboratory (NREL), 2013. URL `https://www.nrel.gov/docs/fy13osti/56723.pdf`.

S. A. Malcolm and S. A. Zenios. Robust optimization for power systems capacity expansion under uncertainty. *The Journal of the Operational Research Society*, 45(9):1040–1049, 1994. ISSN 01605682, 14769360. URL `http://www.jstor.org/stable/2584145`.

D. S. Mallapragada, D. J. Papageorgiou, A. Venkatesh, C. L. Lara, and I. E. Grossmann. Impact of model resolution on scenario outcomes for electricity sector system expansion. *Energy*, 163:1231 – 1244, 2018. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2018.08.015.

N. Mann, C.-H. Tsai, G. Gulen, E. Schneider, P. Cuevas, J. B. J. Dyer, T. Zhang, R. Baldick, T. Deetjen, and R. Morneau. Capacity Expansion and Dispatch Modeling: Model Documentation and Results for ERCOT Scenarios. Technical report, The University of Texas at Austin, Austin, TX, 2017. URL `https://energy.utexas.edu/sites/default/files/UTAustin_FCe_ERCOT_2017.pdf`.

G. P. McCormick. Computability of global solutions to factorable nonconvex programs: Part i — convex underestimating problems. *Mathematical Programming*, 10(1):147–175, 1976. ISSN 1436-4646. doi: 10.1007/BF01580665.

D. Mejia-Giraldo. *Robust and flexible planning of power system generation capacity*. PhD thesis, Iowa State University, 2013.

D. Mejía-Giraldo and J. D. McCalley. Maximizing future flexibility in electric generation portfolios. *IEEE Transactions on Power Systems*, 29(1):279–288, Jan 2014. ISSN 0885-8950. doi: 10.1109/TPWRS.2013.2280840.

R. Misener and C. A. Floudas. Antigone: Algorithms for continuous / integer global optimization of nonlinear equations. *J. of Global Optimization*, 59(2-3):503–526, July 2014. ISSN 0925-5001. doi: 10.1007/s10898-014-0166-2.

R. Misener, J. P. Thompson, and C. A. Floudas. Apogee: Global optimization of standard, generalized, and extended pooling problems via linear and logarithmic partitioning schemes. *Computers & Chemical Engineering*, 35(5):876 – 892, 2011. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2011.01.026. URL `http://www.sciencedirect.com/science/article/pii/S0098135411000366`. Selected Papers from ESCAPE-20 (European Symposium of Computer Aided Process Engineering - 20), 6-9 June 2010, Ischia, Italy.

M. Mistry, A. C. D'Iddio, M. Huth, and R. Misener. Satisfiability modulo theories for process systems engineering. *Computers & Chemical Engineering*, 113:98 – 114, 2018.

ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2018.03.004. URL http://www.sciencedirect.com/science/article/pii/S0098135418301303.

A. Moreira, D. Pozo, A. Street, and E. Sauma. Reliable renewable generation and transmission expansion planning: Co-optimizing system's resources for meeting renewable targets. *IEEE Transactions on Power Systems*, 32(4):3246–3257, July 2017. ISSN 0885-8950. doi: 10.1109/TPWRS.2016.2631450.

J. A. Moulijn, A. Stankiewicz, J. Grievink, and A. Górak. Process intensification and process systems engineering: A friendly symbiosis. *Computers & Chemical Engineering*, 32(1):3 – 11, 2008. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2007.05.014. URL http://www.sciencedirect.com/science/article/pii/S0098135407001512. Process Systems Engineering: Contributions on the State-of-the-Art.

J. M. Mulvey and A. Ruszczyński. A new scenario decomposition method for large-scale stochastic optimization. *Operations Research*, 43(3):477–490, 1995. doi: 10.1287/opre.43.3.477.

F. D. Munoz and J.-P. Watson. A scalable solution framework for stochastic transmission and generation planning problems. *Computational Management Science*, 12(4):491–518, Oct 2015. ISSN 1619-6988. doi: 10.1007/s10287-015-0229-y.

F. D. Munoz, B. F. Hobbs, and S. Kasina. Efficient proactive transmission planning to accommodate renewables. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–7, July 2012. doi: 10.1109/PESGM.2012.6345237.

F. D. Munoz, E. E. Sauma, and B. F. Hobbs. Approximations in power transmission planning: implications for the cost and performance of renewable portfolio standards. *Journal of Regulatory Economics*, 43(3):305–338, Jun 2013. ISSN 1573-0468. doi: 10.1007/s11149-013-9209-8. URL https://doi.org/10.1007/s11149-013-9209-8.

M. Muratori, K. Calvin, M. Wise, P. Kyle, and J. Edmonds. Global economic consequences of deploying bioenergy with carbon capture and storage (BECCS). *Environmental Research*

Letters, 11(9):095004, aug 2016. doi: 10.1088/1748-9326/11/9/095004. URL https://doi.org/10.1088%2F1748-9326%2F11%2F9%2F095004.

P. Nahmmacher, E. Schmid, L. Hirth, and B. Knopf. Carpe diem: A novel approach to select representative days for long-term power system modeling. *Energy*, 112:430 – 442, 2016a. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2016.06.081. URL http://www.sciencedirect.com/science/article/pii/S0360544216308556.

P. Nahmmacher, E. Schmid, L. Hirth, and B. Knopf. Carpe diem: A novel approach to select representative days for long-term power system modeling. *Energy*, 112:430 – 442, 2016b. ISSN 0360-5442. doi: http://doi.org/10.1016/j.energy.2016.06.081. URL http://www.sciencedirect.com/science/article/pii/S0360544216308556.

NERC. Accommodating High Levels of Variable Generation. Technical Report April, North American Electric Reliability Corporation, 2009. URL http://www.nerc.com/files/ivgtf_report_041609.pdf.

S. A. Newell, K. Spees, J. P. Pfeifenberger, I. Karkatsou, N. Wintermantel, and K. Carden. Estimating the economically optimal reserve margin in ERCOT. Technical report, The Brattle Group and Astrape Consulting, Cambridge, MA, 2014. URL http://www.brattle.com/system/news/pdfs/000/000/613/original/Estimating_the_Economically_Optimal_Reserve_Margin_in_ERCOT.pdf?1391445083.

N. Noyan. Risk-averse two-stage stochastic programming with an application to disaster management. *Computers & Operations Research*, 39(3):541 – 559, 2012. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2011.03.017. URL http://www.sciencedirect.com/science/article/pii/S0305054811000931.

NREL. Annual Technology Baseline and Standard Scenarios. http://www.nrel.gov/analysis/docs/ATB_Data_Inputs_V6.xlsm, 2016. Online; Accessed 7 September 2016.

NWPCC. Sixth Northwest Conservation and Electric Power Plan. Technical Report February, Northwest Power and Conservation Council, 2010. URL `https://www.nwcouncil.org/energy/powerplan/6/plan/`.

R. P. O'Neill, E. A. Krall, K. W. Hedman, and S. S. Oren. A model and approach to the challenge posed by optimal power systems planning. *Mathematical Programming*, 140(2):239–266, 2013. ISSN 1436-4646. doi: 10.1007/s10107-013-0695-3. URL `http://dx.doi.org/10.1007/s10107-013-0695-3`.

V. Oree, S. Sayed Hassen, and P. Fleming. Generation expansion planning optimisation with renewable energy integration: A review. *Renewable and Sustainable Energy Reviews*, 69: 790–803, 2017. doi: 10.1016/j.rser.2016.11.120.

D. Ozokwelu. High Efficiency Modular Chemical Processes (HEMCP). Technical report, U.S. Department of Energy, September 2014. URL `https://www.energy.gov/sites/prod/files/2014/10/f18/hemcp-topic-overview.pdf`.

B. Palmintier and M. Webster. Impact of unit commitment constraints on generation expansion planning with renewables. In *2011 IEEE Power and Energy Society General Meeting*, pages 1–7, July 2011. doi: 10.1109/PES.2011.6038963.

B. S. Palmintier and M. D. Webster. Heterogeneous unit clustering for efficient operational flexibility modeling. *IEEE Transactions on Power Systems*, 29(3):1089–1098, May 2014. ISSN 0885-8950. doi: 10.1109/TPWRS.2013.2293127.

H. Park and R. Baldick. Transmission planning under uncertainties of wind and load: Sequential approximation approach. *IEEE Transactions on Power Systems*, 28(3):2395–2402, Aug 2013. ISSN 0885-8950. doi: 10.1109/TPWRS.2013.2251481.

H. Park and R. Baldick. Multi-year stochastic generation capacity expansion planning under environmental energy policy. *Applied Energy*, 183:737 – 745, 2016. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2016.08.164. URL `http://www.sciencedirect.com/science/article/pii/S0306261916312739`.

M. V. F. Pereira and L. M. V. G. Pinto. Stochastic optimization of a multireservoir hydroelectric system: A decomposition approach. *Water Resources Research*, 21(6):779–792, 1985. doi: 10.1029/WR021i006p00779. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR021i006p00779.

M. V. F. Pereira and L. M. V. G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1):359–375, May 1991. ISSN 1436-4646. doi: 10.1007/BF01582895. URL https://doi.org/10.1007/BF01582895.

J. P. Pfeifenberger, K. Spees, K. Carden, and N. Wintermantel. Resource Adequacy Requirements: Reliability and Economic Implications. Technical report, The Brattle Group and Astrape Consulting, Cambridge, MA, 2013. URL https://www.ferc.gov/legal/staff-reports/2014/02-07-14-consultant-report.pdf.

A. Philpott and Z. Guan. On the convergence of stochastic dual dynamic programming and related methods. *Operations Research Letters*, 36(4):450 – 455, 2008. ISSN 0167-6377. doi: https://doi.org/10.1016/j.orl.2008.01.013.

A. Pina, C. Silva, and P. Ferrão. Modeling hourly electricity dynamics for policy making in long-term scenarios. *Energy Policy*, 39(9):4692 – 4702, 2011. ISSN 0301-4215. doi: http://doi.org/10.1016/j.enpol.2011.06.062. URL http://www.sciencedirect.com/science/article/pii/S0301421511005180.

A. Pina, C. A. Silva, and P. Ferrão. High-resolution modeling framework for planning electricity systems with high penetration of renewables. *Applied Energy*, 112:215 – 223, 2013. ISSN 0306-2619. doi: http://dx.doi.org/10.1016/j.apenergy.2013.05.074. URL http://www.sciencedirect.com/science/article/pii/S030626191300487X.

R. J. Pinto, C. T. Borges, and M. E. P. Maceira. An efficient parallel algorithm for large scale hydrothermal system operation planning. *IEEE Transactions on Power Systems*, 28 (4):4888–4896, Nov 2013. ISSN 0885-8950. doi: 10.1109/TPWRS.2012.2236654.

K. Poncelet, E. Delarue, J. Duerinck, D. Six, and W. D'haeseleer. The importance of integrating the variability of renewables in long-term energy planning models. *KU Leuven*, pages 1–18, October 2014. URL `https://www.mech.kuleuven.be/en/tme/research/energy_environment/Pdf/wp-importance.pdf`.

K. Poncelet, H. Hoschle, E. Delarue, and W. D'haeseleer. Selecting representative days for investment planning models. *KU Leuven*, 2015. URL `https://www.mech.kuleuven.be/en/tme/research/energy_environment/Pdf/wpen201510.pdf`.

Potomac Economics. 2015 State of the Market Report for the ERCOT Whole Electricity Markets. Technical Report June, Potomac Economics, 2016. URL `http://www.puc.texas.gov/industry/electric/reports/ERCOT_annual_reports/2015annualreport.pdf`.

W. B. Powell. *Clearing the Jungle of Stochastic Optimization*, pages 109–137. 09 2014. ISBN 978-0-9843378-5-9. doi: 10.1287/educ.2014.0128.

D. Pozo, E. E. Sauma, and J. Contreras. A three-level static milp model for generation and transmission expansion planning. *IEEE Transactions on Power Systems*, 28(1):202–210, Feb 2013. ISSN 0885-8950. doi: 10.1109/TPWRS.2012.2204073.

D. Pozo, J. Contreras, and E. E. Sauma. Unit commitment with ideal and generic energy storage units. *IEEE Transactions on Power Systems*, 29(6):2974–2984, Nov 2014. ISSN 0885-8950. doi: 10.1109/TPWRS.2014.2313513.

S. T. Rachev and W. Römisch. Quantitative stability in stochastic programming: The method of probability metrics. *Mathematics of Operations Research*, 27(4):792–818, 2002. ISSN 0364765X, 15265471. URL `http://www.jstor.org/stable/3690468`.

R. Rahmaniani, T. G. Crainic, M. Gendreau, and W. Rei. The Benders decomposition algorithm: A literature review. *European Journal of Operational Research*, 259(3):801–817, 2016. ISSN 0377-2217. doi: http://dx.doi.org/10.1016/j.ejor.2016.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0377221716310244`.

S. Rebennack. Generation expansion planning under uncertainty with emissions quotas. *Electric Power Systems Research*, 114:78 – 85, 2014. ISSN 0378-7796. doi: https://doi.org/10.1016/j.epsr.2014.04.010.

S. Rebennack. Combining sampling-based and scenario-based nested benders decomposition methods: Application to stochastic dual dynamic programming. *Math. Program.*, 156 (1-2):343–389, Mar. 2016. ISSN 0025-5610. doi: 10.1007/s10107-015-0884-3.

S. Rebennack, B. Flach, M. V. F. Pereira, and P. M. Pardalos. Stochastic hydro-thermal scheduling underco_2emissions constraints. *IEEE Transactions on Power Systems*, 27(1): 58–68, Feb 2012. ISSN 0885-8950. doi: 10.1109/TPWRS.2011.2140342.

G. G. Rogers and L. Bottaci. Modular production systems: a new manufacturing paradigm. *Journal of Intelligent Manufacturing*, 8(2):147–156, Mar 1997. ISSN 1572-8145. doi: 10.1023/A:1018560922013. URL `https://doi.org/10.1023/A:1018560922013`.

J. H. Roh, M. Shahidehpour, and Y. Fu. Market-based coordination of transmission and generation capacity planning. *IEEE Transactions on Power Systems*, 22(4):1406–1419, Nov 2007. ISSN 0885-8950. doi: 10.1109/TPWRS.2007.907894.

S. Roy. Consider modular plant design. *Chemical Engineering Progress*, (113):28–31, 2017.

H. Sadeghi, M. Rashidinejad, and A. Abdollahi. A comprehensive sequential review study through the generation expansion planning. *Renewable and Sustainable Energy Reviews*, 67:1369–1394, 2017. doi: 10.1016/j.rser.2016.09.046.

N. Sahinidis and I. Grossmann. Convergence properties of generalized benders decomposition. *Computers and Chemical Engineering*, 15(7):481 – 491, 1991. ISSN 0098-1354. doi: http://dx.doi.org/10.1016/0098-1354(91)85027-R. URL `http://www.sciencedirect.com/science/article/pii/009813549185027R`.

N. Sawaya. *Reformulations, relaxations and cutting planes for generalized disjunctive programming*. PhD thesis, Carnegie Mellon University, 2006.

O. Schmidt, A. Hawkes, A. Gambhir, and I. Staffell. The future cost of electrical energy storage based on experience rates. *Nature Energy*, 2(17110):1–1, 2017. doi: 10.1038/nenergy.2017.110.

A. Shapiro, W. Tekaya, J. P. d. Costa, and M. P. Soares. Risk neutral and risk averse stochastic dual dynamic programming method. *European Journal of Operational Research*, 224(2):375 – 391, 2013a. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2012.08.022.

A. Shapiro, W. Tekaya, J. P. da Costa, and M. P. Soares. Risk neutral and risk averse stochastic dual dynamic programming method. *European Journal of Operational Research*, 224(2):375 – 391, 2013b. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2012.08.022. URL `http://www.sciencedirect.com/science/article/pii/S0377221712006455`.

A. D. Sherali and C. M. Shetty. The rectilinear distance location-allocation problem. *A I I E Transactions*, 9(2):136–143, 1977. doi: 10.1080/05695557708975135.

H. D. Sherali and F. L. Nordai. NP-Hard, capacitated, balanced p-median problems on a chain graph with a continuum of link demands. *Mathematics of Operations Research*, 13 (1):32–49, 1988. ISSN 0364-765X. doi: 10.1287/moor.13.1.32.

H. D. Sherali and C. H. Tuncbilek. A squared-euclidean distance location-allocation problem. *Naval Research Logistics (NRL)*, 39(4):447–469, 1992. ISSN 1520-6750. doi: 10.1002/1520-6750(199206)39:4<447::AID-NAV3220390403>3.0.CO;2-O.

H. D. Sherali, I. Al-Loughani, and S. Subramanian. Global optimization procedures for the capacitated euclidean and ℓ_p distance multifacility location-allocation problems. *Operations Research*, 50(3):433–448, 2002. ISSN 0030-364X. doi: 10.1287/opre.50.3.433.7739.

W. Short, P. Sullivan, T. Mai, M. Mowers, C. Uriarte, N. Blair, D. Heimiller, and A. Martinez. Regional Energy Deployment System (ReEDS). Technical Report December, National Renewable Technology Laboratory (NREL), 2011. URL `https://www.nrel.gov/analysis/reeds/pdfs/reeds_documentation.pdf`.

A. Shortt and M. O'Malley. Impact of variable generation in generation resource planning models. In *IEEE PES General Meeting*, pages 1–6, July 2010. doi: 10.1109/PES.2010.5589461.

G. Steeger and S. Rebennack. Dynamic convexification within nested benders decomposition using lagrangian relaxation: An application to the strategic bidding problem. *European Journal of Operational Research*, 257(2):669 – 686, 2017. ISSN 0377-2217. doi: http://dx.doi.org/10.1016/j.ejor.2016.08.006. URL `http://www.sciencedirect.com/science/article/pii/S037722171630621X`.

M. Tawarmalani and N. V. Sahinidis. A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming*, 103(2):225–249, 2005. ISSN 00255610. doi: 10.1007/s10107-005-0581-8.

F. Thome, M. Pereira, S. Granville, and M. Fampa. Non-Convexities Representation on Hydrothermal Operation Planning using SDDP. *Submitted for publication*, 2013. URL `http://www.psr-inc.com/publications`.

R. Tidball, J. Bluestein, N. Rodriguez, and S. Knoke. Cost and Performance Assumptions for Modeling Electricity Generation Technologies," National Renewable Energy Laboratory. Technical report, National Renewable Energy Laboratory, Golden, CO, 2010. URL `https://www.nrel.gov/docs/fy11osti/48595.pdf`.

F. Trespalacios. *Improved formulations and computational strategies for the solution of convex and nonconvex generalized disjunctive programs*. PhD thesis, Carnegie Mellon University, 2015. URL `https://search.proquest.com/docview/1763825553?accountid=9902`.

F. Trespalacios and I. E. Grossmann. Review of mixed-integer nonlinear and generalized disjunctive programming methods. *Chemie-Ingenieur-Technik*, 86(7):991–1012, 2014. ISSN 15222640. doi: 10.1002/cite.201400037.

J. Truby. Thermal Power Plant Economics and Variable Renewble Energies. Technical report, International Energy Agency, Paris, France, 2014. URL https://www.iea.org/publications/insights/insightpublications/thermal-power-plant-economics-and-variable-renewable-energies.html.

Union of Concerned Scientists. The Nuclear Power Dilemma. Technical report, Union of Concerned Scientists, 2018. URL https://www.ucsusa.org/sites/default/files/attach/2018/11/Nuclear-Power-Dilemma-executive-summary.pdf.

J. M. Velásquez Bermúdez. Gddp: Generalized dual dynamic programming theory. *Annals of Operations Research*, 117(1):21–31, Nov 2002. ISSN 1572-9338. doi: 10.1023/A:1021557003554. URL https://doi.org/10.1023/A:1021557003554.

J. P. Vielma and G. L. Nemhauser. Modeling disjunctive constraints with a logarithmic number of binary variables and constraints. *Mathematical Programming*, 128(1):49–72, Jun 2011. ISSN 1436-4646. doi: 10.1007/s10107-009-0295-4. URL https://doi.org/10.1007/s10107-009-0295-4.

J.-P. Watson and D. L. Woodruff. Progressive hedging innovations for a class of stochastic mixed-integer resource allocation problems. *Computational Management Science*, 8(4):355–370, Nov 2011. ISSN 1619-6988. doi: 10.1007/s10287-010-0125-4. URL https://doi.org/10.1007/s10287-010-0125-4.

A. Weber and C. J. Friedrich. *Theory of the Location of Industries*. University of Chicago Press, Chicago, IL, 1929.

M. Webster, P. Donohoo, and B. Palmintier. Water–co2 trade-offs in electricity generation planning. *Nature Climate Change*, 3:1029–1032, 2013. URL https://doi.org/10.1038/nclimate2032.

WEC. The Westinghouse Pressurized Water Reactor Nuclear Power Plant. Technical report, Westinghouse Electric Corporation, 1984. URL http://www4.ncsu.edu/~doster/NE405/Manuals/PWR_Manual.pdf.

H. P. Williams. *Model building in mathematical programming*. John Wiley & Sons, 2013.

S. Wogrin, E. Centeno, and J. Barquin. Generation capacity expansion in liberalized electricity markets: A stochastic mpec approach. *IEEE Transactions on Power Systems*, 26 (4):2526–2532, Nov 2011. ISSN 0885-8950. doi: 10.1109/TPWRS.2011.2138728.

L. A. Wolsey and G. L. Nemhauser. *Integer and combinatorial optimization*. John Wiley & Sons, 2014.

F. Wood. The Consideration of PV Curtailments in NEMS: Addressing the Duck Problem. `http://www.eia.gov/renewable/workshop/pdf/session2_Wood.pdf`, 2016. Online; Accessed 10 October 2016.

F. You and B. Wang. Life cycle optimization of biomass-to-liquid supply chains with distributed–centralized processing networks. *Industrial & Engineering Chemistry Research*, 50(17):10102–10127, 2011. doi: 10.1021/ie200850t. URL `https://doi.org/10.1021/ie200850t`.

Y. Zhan, Q. P. Zheng, J. Wang, and P. Pinson. Generation expansion planning with large amounts of wind power via decision-dependent stochastic programming. *IEEE Transactions on Power Systems*, 32(4):3015–3026, July 2017. ISSN 0885-8950. doi: 10.1109/TPWRS.2016.2626958.

Q. Zhang, I. E. Grossmann, and R. M. Lima. On the relation between flexibility analysis and robust optimization for linear systems. *AIChE Journal*, 62(9):3109–3123, 2016. doi: 10.1002/aic.15221.

J. Zhu and M.-Y. Chow. A review of emerging techniques on generation expansion planning. *IEEE Transactions on Power Systems*, 12(4):1722–1728, Nov 1997. ISSN 0885-8950. doi: 10.1109/59.627882.

J. Zou, S. Ahmed, and X. A. Sun. Multistage Stochastic Unit Commitment Using Stochastic Dual Dynamic Integer Programming. *Submitted for publication*, 2017. URL `http://www.optimization-online.org/DB_HTML/2017/05/6003.html`.

J. Zou, S. Ahmed, and X. A. Sun. Partially adaptive stochastic optimization for electric power generation expansion planning. *INFORMS Journal on Computing*, 30(2):388–401, 2018a. doi: 10.1287/ijoc.2017.0782.

J. Zou, S. Ahmed, and X. A. Sun. Stochastic dual dynamic integer programming. *Mathematical Programming*, Mar 2018b. ISSN 1436-4646. doi: 10.1007/s10107-018-1249-5.

Appendix A

Chapter 4 additional material

A.1 Detailed Algebraic Modeling Descriptions

This appendix provides the detailed mathematical optimization formulations associated with the chronological capacity expansion model (C-GEP) and the time slice capacity expansion model (TS-GEP). Regarding their similarities, both are deterministic optimization models that take the vantage point of a centralized planner seeking to determine cost-optimal expansion decisions over a planning horizon of several decades. Both models build and utilize generation capacity to satisfy load in their respective time steps. Both models use as input the same forecasted load growth, the same suite of generation technologies to meet this growth, and the same associated cost assumptions to model grid evolution in 3-year time increments from 2015 to 2045. They represent solar and wind expansion decisions as "continuous" decisions meaning that a fractional wind generator can be built. Importantly, both models represent the existing fleet of thermal and renewable generators (wind and solar) by clustering the entire fleet into seven different generator types. C-GEP and TS-GEP also allow for aging capacity to be retired or retrofitted, whereby the latter options incurs a one-time cost of retrofit and returns to operation with the same operational parameters as before. For each generation technology, both models include annual capacity installation limits that implicitly account for supply chain constraints associated with technology deployment.

Despite these similarities, the two models significantly differ in their temporal resolution and operational detail. It is precisely the dissimilarities described below that will help elucidate why different expansion decisions are made in certain scenarios. Temporally, the C-GEP represents monthly load, as well as wind and solar generation, by a single day at an hourly time resolution, whereas the TS-GEP represents annual load, as well as wind and solar generation, with 16 time slices representing different times of day and seasons. In particular, while the chronological model "sees" an hourly load and renewables capacity factor time series corresponding to twelve representative days of the year. In other words, the TS-GEP averages load and renewables capacity factor data in each of the four seasons into time slices representing morning (7 am - 2 pm), afternoon (2-6 pm), evening (6-11 pm), and night (11 pm - 7 am). More importantly, the C-GEP, as its name suggests, "sees" chronology, and therefore events that occur in a given hour are related to events that occur in the preceding and subsequent hours. The TS-GEP does not link two consecutive time slices with respect to operational constraints. Operationally, C-GEP considers important details associated with thermal generators including: unit commitment decisions, ramping constraints, spinning reserves, quick-start reserves, and start-up costs. In contrast, the TS-GEP omits these details, although spinning reserves are partially taken into account. Lastly, because the C-GEP includes unit commitment decisions, thermal generation expansion decisions are modeled as integer decisions, unlike the TS-GEP which allows for a fractional number of thermal generators to be built.

Both models were implemented in GAMS. The descriptions below were generated using the GAMS utility function `model2tex` as described on the GAMS website. Table A.1 lists the section references for key constraints common to both models.

Constraint type	**Chronological C-GEP**	**Time Slice TS-GEP**
Load balance	A.1.1	A.1.2
Generator capacity balance	A.1.1–A.1.1	A.1.2
Retirement or retrofit decisions	A.1.1–A.1.1	A.1.2
Annual installation limits	A.1.1–A.1.1	A.1.2
RPS	A.1.1	A.1.2
Capacity planning reserve requirements	A.1.1	A.1.2
System operating reserve requirements	A.1.1	A.1.2
Spinning reserves for thermal generation	A.1.1	A.1.2
Cluster commitment status	A.1.1	NA
Ramping	A.1.1–A.1.1	NA
Power output from generators upper and lower bounds	A.1.1–A.1.1	NA
Minimum turndown	NA	A.1.2

Table A.1: Section references to specific constraints

A.1.1 Chronological Capacity Expansion Model (C-GEP)

Sets

Name	**Domains**	**Description**
t, tt	t	Set of years to be modeled
h, hh	*	Set of time blocks or hours within each dispatch period
d	*	Set of dispatch periods
s, ss	s	Set of nodes. Model represents ERCOT region with a single node
g, gg	g	Set of generator clusters
Wind	g	Set of wind generators - old and new
CSP	g	Set of CSP generators
PV	g	Set of PV generators
Renew	g	Set of renewable generators
ExistingRenew	g	Set of existing renewable generators
NewRenew	g	Set of new renewable generators

Name	Domains	Description
Thermal	g	Set of thermal generators
Existingthermal	g	Set of existing thermal generators
Newthermal	g	Set of new thermal generators
Thermalbase	g	Set of thermal generators with non-zero minimum output
ThermalQstart	g	Set of thermal generators contributing to quick start reserves
ThermalSpin	g	Set of thermal generators contributing to spinning reserves

Parameters

Name	Domains	Description
CapCost	g, t	Annualized investment cost of generator type g in time period t ($ per kW)
LifeExtensionCost	g, t	One time extension cost to extend plant beyond its economic life ($ per kW)
Investmult	g, t	Portion of overnight capital cost of generator type g in time period t that is included in objective
Capmult	g	Technology-specific financial multiplier to account for any applicable differences in depreciation schedule, and tax policies for each generator g >1
FOMCost	g, t	Annual fixed operating & maintenance costs for generator g ($ per kW-year)
VOMCost	g, t	Variable operating & maintenance costs for generator g ($ per MWh)

Name	**Domains**	**Description**
FuelCost	g, t	Fuel costs for generator type g in time period t (\$ per MMBtu)
StartupFueluse	g	Fuel use during startup for each generator type g (MMBtu per MW)
StartUpcost	g, t	Startup cost of generator type g in time period t (\$ per MW)
Gridconnect	g, t	Annualized cost of connecting a new generator of cluster g to the grid (\$ per kW)
CarbonTax	t	Carbon tax on emissions from power plants during year t (\$ per ton CO2eq)
Ngenexist	g, s	Number of units for each generator cluster at node s at t=0
PTCEligiblePlants	g, s, t	Number of existing wind plants that are eligible for the production tax credit in each year t
Cf	g, s, h, d	Capacity factor for generator type g for each time instance
Heatrate	g, s	Heat rate of generator type g (Btu per kWh)
Pgen	g, s	Assumed size or capacity of an individual type g generator in node s (MW)
Pgenmin	g	Minimum operating capacity of an individual type g generator (MW)
Tlife	g	Lifetime of generator type g (years)
DPdown	g	Maximum ramp down rate for thermal generator between two consecutive time blocks (% of nameplate capacity per hour)

Name	Domains	Description
DPup	g	Maximum ramp up rate for thermal generator between two consecutive time blocks (% of nameplate capacity per hour)
Spinfrac	g	Fraction of nameplate capacity that can contribute to spinning reserves for generator type g
Qstartfrac	g	Fraction of nameplate capacity that can contribute to quick start reserves for generator type g
Nmaxgeninstall	g, t	Maximum no. of generators of type g that can be installed at beginning of each year
Emissionf	g	Greenhouse gas (GHG) emissions factor of fuel used by generation type g (kg CO2eq per MMBtu fuel)
CV	g, t	Capacity value or fraction of installed capacity of generator g contributing to planning reserve requirement
Load	s, h, d, t	Demand at node s in block h of dispatch period d in year t (MW)
MaxLoadMW	t	Maximum load to define planning reserve requirement(MW)
Curtailcost		Cost of curtailing generation in $ per MWh
Seasonscale	d	Weight to scale generation in each hour of representative day to d to its contribution to annual generation
RPSfrac	t	Fraction of annual load met by renewables in year t

Name	Domains	Description
RPSCapacityMin	t	Minimum renewable generating capacity to exist in each year in MW - included to modeled Texas renewables policy
WindPTC	t, tt	Production tax credit (PTC) for electricity produced from wind plants in period t that were built in periodd tt ($ per MWh) - applies to only first 10 years of operation of each plant
PTCforExistingPlants		PTC credit for each existing wind plant - $23 per MWh
Reservemargin		Reserve margin for planning reserves (% of load)
ORspin	h, d	Operating spinning reserves for each block h in dispatch period d (%)
ORTot	h, d	Total operating reserve requirement (spinning + non-spinning) for each block h in dispatch period d (%)
DiscountFact		Scalar (between 0 and 1) used to discount future year costs in the objective function to year 1 that allows consideration of inter-temporal trade-offs in the model
Dur	h	Duration of time block h = 1 hour
Yearrel	t	Years considered in the model (relative to t=0)
Opexmult	t	Multiplier to account for discounted operating cost for intermediate years between planning years
Objval_Units		Units in which the objective function value is measured (e.g. 1e9 ==>US$1 billion)

Variables

Name	**Domains**	**Description**
NGenR	g, s, t	No. of renewable generators of type g that are available at node s at beginning of year t
NGenInstallR	g, s, t	No. of renewable generators of type g that are installed at node s at beginning of year t
NGenRetireR	g, s, t	No. of renewable generators of type g that are retired at node s at beginning of year t
NGenRExtend	g, s, t	Number of generators of type g at node s that are extended beyond their economic lifetime at beginning of year t
NGenRInstalledCapTotal	g, s, t	Total number of generators of type g at node s that have not been retired at beginning of year t
AvgPower	g, s, h, d, t	Average power from generator g at node s in block h of dispatch period d in year t (MW)
SpinCap	g, s, h, d, t	Spinning capacity of generator type g located at node s and reserved during block h of dispatch period d in year t (MW))
Curtail	s, h, d, t	Total curtailed generation at each node s during each hour h dispatch period d and year t (MW)
Qstartcap	g, s, h, d, t	Quick start capacity of generator type g located at node s and reserved during block h of dispatch period d in year t (MW))
ObjInvgenTcost	t	Cost of installing new thermal generators in year t (billion $)
ObjInvgenRcost	t	Cost of installing new renewable generators in year t (billion $)

Name	**Domains**	**Description**
ObjRetrofitcost	t	Cost of extending lifetime of existing generator in year t (billion $)
ObjFOMTcost	t	Fixed operating and maintenance (FOM) cost of thermal generators in year t (billion $)
ObjFOMRcost	t	Fixed operating and maintenance cost of renewable generators in year t (billion $)
ObjVOMTcost	t	Variable operating and maintenance (VOM) cost of thermal generators in year t, excluding fuel costs (billion $)
ObjFuelTcost	t	Fuel costs of thermal generators in year t (billion $)
ObjEnvgencost	t	Environmental policy costs of generation in year t (billion $)
ObjStartupcost	t	Startup cost of thermal generators in year t (billion $)
ObjCurtailcost	t	Cost of curtailment for each year t
ObjVOMRcost	t	Variable operating cost of renewable generators in year t, (billion $)
ObjCredits	t	Production and investment tax credits for renewable generation in year t
NGenT	g, s, t	Integer number of generators of type g at node s that are before their economic lifetime and at beginning of year t
NGenTExtend	g, s, t	Integer number of thermal generators of type g at node s that are extended beyond their economic lifetime at beginning of year t

Name	Domains	Description
NGenTInstalledCapTotal	g, s, t	Integer number of thermal generators of type g at node s that have not been retired at beginning of year t
NGenInstallT	g, s, t	Integer number of thermal generators of type g at node s installed at beginning of year t
NGenRetireT	g, s, t	Integer number of thermal generators of type g at node s that are retired at beginning of year t
NShutD	g, s, h, d, t	Integer number of generators of type g at node s that are shutdown at beginning of block h of period d in year t
NStartUp	g, s, h, d, t	Integer number of generators of type g at node s that are started at beginning of block h of period d in year t
NGenOn	g, s, h, d, t	Integer number of generators of type g at node s that are on during block h of period d in year t
objval		Objective function value

Equations

Name	Domains	Description
ObjectiveFunction		Objective function minimizing total annualized cost
Loadbal	s, h, d, t	Load balance at each node and time instance
NGenOndef	g, s, h, d, t	Defining number of generators that are turned on
Rampdownlim	g, s, h, d, t	Defining Ramp down limits for thermal generators

Name	Domains	Description
Rampuplim	g, s, h, d, t	Defining Ramp up limits for thermal generators
Gencapexistdef	g, s, t	Tracking total existing thermal generation capacity
Gencapnewdef	g, s, t	Tracking total new thermal generation capacity
NewThermalendoflife	g, s, t	Determining end of lifetime state for new thermal generators
NGenTInstalledCapTotaldef	g, s, t	Total thermal generation capacity either in extended condition or pre-retirement condition
Renewcapexistdef	g, s, t	Tracking total existing renewable generation capacity
Renewcapnewdef	g, s, t	Tracking total new renewable generation capacity
Newrenewendoflife	g, s, t	Determining end of lifetime state for new renewable generators
NGenRInstalledCapTotaldef	g, s, t	Total renewable generation capacity either in extended condition or pre-retirement condition
Planningreservedef	t	Planning reserve requirement for the entire region for each year t
TotalOpreservedef	h, d, t	Total operating reserve requirement for the entire region for each time instance
Spinningreservedef	h, d, t	Spinning reserve requirement for the entire region for each time instance
Qstartfracdef	g, s, h, d, t	Fraction of generation capacity allocated to quick start reserves
Spinfracdef	g, s, h, d, t	Fraction of generation capacity allocated to spinning reserves

Name	Domains	Description
OpcapUB	g, s, h, d, t	Operating capacity upper bound for thermal generators
OpcapLB	g, s, h, d, t	Lower bound on operating capacity for thermal generators when ON
RenewoutputUB	g, s, h, d, t	Upper bound on renewable energy output for each hour
NGenOnUB	g, s, h, d, t	No. of generators turned on at any period cannot exceed total number of installed generators
NGenInstallTUB	g, t	Upper bound on number of thermal generators installed of type g in a year based on annual installation limits
NGenInstallRUB	g, t	Upper bound on number of renewable generators installed of type g in a year
RPSconstr	t	Minimum renewable penetration constraint for entire region
RPScapacityconstr	t	Constraint on minimum generating capacity of renewable resources as per policy requirements for Texas region
ObjInvgenTcostdef	t	Defining investment cost of thermal generators
ObjInvgenRcostdef	t	Defining investment cost of renewable generators
ObjRetrofitcostdef	t	Cost of extending lifetime of existing thermal generators
ObjFOMTcostdef	t	FOM cost of thermal generators
ObjFOMRcostdef	t	FOM cost of renewable generators
ObjVOMTcostdef	t	VOM cost of thermal generators, excluding fuel costs

Name	**Domains**	**Description**
ObjFuelTcostdef	t	Fuel cost of thermal generators
ObjVOMRcostdef	t	VOM cost of renewable generators -excludes fuel
ObjEnvgencostdef	t	Environmental policy cost of generation
ObjStartupcostdef	t	Startup cost and shutdown costs of thermal generators
Objcurtailcostdef	t	Cost penalty for curtailing renewable generation
ObjCreditsdef	t	Production and investment tax credits for renewable generation

Equation Definitions

ObjectiveFunction:

$$\text{objval} = \sum_t \Big(\frac{1}{(1 + \text{DiscountFact})^{(\text{Yearrel}_t - 1)}} \cdot \big(\text{ObjInvgenTcost}_t + \text{ObjInvgenRcost}_t + \text{ObjRetrofitcost}_t + \text{ObjFOMTcost}_t + \text{ObjFOMRcost}_t + \text{ObjVOMTcost}_t + \text{ObjFuelTcost}_t + \text{ObjVOMRcost}_t + \text{ObjEnvgencost}_t + \text{ObjStartupcost}_t + \text{ObjCurtailcost}_t - \text{ObjCredits}_t \big) \Big)$$

Loadbal$_{s,h,d,t}$**:**

$$\sum_g (\text{AvgPower}_{g,s,h,d,t}) = \text{Load}_{s,h,d,t} + \text{Curtail}_{s,h,d,t} \qquad \forall s, h, d, t$$

NGenOndef$_{g,s,h,d,t}$**:**

$$\text{NGenOn}_{g,s,h,d,t} = \text{NGenOn}_{g,s,h--1,d,t} + \text{NStartUp}_{g,s,h,d,t} - \text{NShutD}_{g,s,h,d,t} \quad \forall g, s, h, d, t \mid \text{Thermalbase}_g$$

Rampdownlim$_{g,s,h,d,t}$**:**

$$\text{AvgPower}_{g,s,h--1,d,t} - \text{AvgPower}_{g,s,h,d,t} \leq (\text{NGenOn}_{g,s,h,d,t} - \text{NStartUp}_{g,s,h,d,t}) \cdot \text{DPdown}_g \cdot \text{Pgen}_{g,s} \cdot \text{Dur}_h + \max((\text{DPdown}_g \cdot \text{Dur}_h), \text{Pgenmin}_g) \cdot \text{Pgen}_{g,s} \cdot \text{NShutD}_{g,s,h,d,t} - \text{Pgenmin}_g \cdot \text{Pgen}_{g,s} \cdot \text{NStartUp}_{g,s,h,d,t} \qquad \forall g, s, h, d, t \mid \text{Thermalbase}_g$$

Rampuplim$_{g,s,h,d,t}$:

$\text{AvgPower}_{g,s,h,d,t} - \text{AvgPower}_{g,s,h--1,d,t} \leq (\text{NGenOn}_{g,s,h,d,t} - \text{NStartUp}_{g,s,h,d,t}) \cdot \text{DPup}_g \cdot \text{Pgen}_{g,s} \cdot \text{Dur}_h + \max((\text{DPup}_g \cdot \text{Dur}_h), \text{Pgenmin}_g) \cdot \text{Pgen}_{g,s} \cdot \text{NStartUp}_{g,s,h,d,t} - \text{Pgenmin}_g \cdot \text{Pgen}_{g,s} \cdot \text{NShutD}_{g,s,h,d,t} \qquad \forall g,s,h,d,t \;|\text{Thermalbase}_g$

Gencapexistdef$_{g,s,t}$:

$\text{NGenT}_{g,s,t} = \text{NGenT}_{g,s,t-1} - \text{NGenRetireT}_{g,s,t} - \text{NGenTExtend}_{g,s,t} + \text{Ngenexist}_{g,s}[(\text{ord(t)} = 1)] \qquad \forall g,s,t \;|\text{Existingthermal}_g$

NGenTInstalledCapTotaldef$_{g,s,t}$:

$$\text{NGenTInstalledCapTotal}_{g,s,t} = \text{NGenT}_{g,s,t} + \sum_{tt|(\text{ord(tt)} \leq \text{ord(t)})} (\text{NGenTExtend}_{g,s,tt}) \;\; \forall g,s,t \;|\text{Thermal}_g$$

Gencapnewdef$_{g,s,t}$:

$\text{NGenT}_{g,s,t} = \text{NGenT}_{g,s,t-1} + \text{NGenInstallT}_{g,s,t} - \text{NGenRetireT}_{g,s,t} \qquad \forall g,s,t \;|\text{Newthermal}_g$

Renewcapexistdef$_{g,s,t}$:

$\text{NGenR}_{g,s,t} = \text{NGenR}_{g,s,t-1} - \text{NGenRetireR}_{g,s,t} - \text{NGenRExtend}_{g,s,t} + \text{Ngenexist}_{g,s}[(\text{ord(t)} = 1)] \qquad \forall g,s,t \;|\text{ExistingRenew}_g$

NGenRInstalledCapTotaldef$_{g,s,t}$:

$$\text{NGenRInstalledCapTotal}_{g,s,t} = \text{NGenR}_{g,s,t} + \sum_{tt|(\text{ord(tt)} \leq \text{ord(t)})} (\text{NGenRExtend}_{g,s,tt}) \qquad \forall g,s,t \;|\text{Renew}_g$$

Renewcapnewdef$_{g,s,t}$:

$\text{NGenR}_{g,s,t} = \text{NGenR}_{g,s,t-1} + \text{NGenInstallR}_{g,s,t} - \text{NGenRetireR}_{g,s,t} - \text{NGenRExtend}_{g,s,t} \;\forall g,s,t \;|\text{NewRenew}_g$

NewThermalendoflife$_{g,s,t}$:

$$\sum_{tt|(\text{ord(tt)} \leq \text{ord(t)})} (\text{NGenRetireT}_{g,s,tt}) \geq \sum_{tt|(\text{Yearrel}_{tt} \leq (\text{Yearrel}_t - \text{Tlife}_g))} (\text{NGenInstallT}_{g,s,tt})$$
$\forall g,s,t \mid ((\text{Tlife}_g \leq (\text{Yearrel}_t - 1)) \wedge \text{Newthermal}_g)$

Newrenewendoflife$_{g,s,t}$:

$$\sum_{tt|(\mathrm{ord(tt)}\leq\mathrm{ord(t)})}\left(\mathrm{NGenRetireR}_{g,s,tt}+\mathrm{NGenRExtend}_{g,s,tt}\right)\geq\sum_{tt|(\mathrm{Yearrel}_{tt}\leq(\mathrm{Yearrel}_t-\mathrm{Tlife}_g))}\left(\mathrm{NGenInstallR}_{g,s,tt}\right)$$
$$\forall g,s,t\mid((\mathrm{Tlife}_g\leq(\mathrm{Yearrel}_t-1))\wedge\mathrm{NewRenew}_g)$$

Planningreservedef$_t$:

$$\sum_{s,Thermal_g}\left(\mathrm{NGenTInstalledCapTotal}_{g,s,t}\cdot\mathrm{Pgen}_{g,s}\right)+\sum_{s,Renew_g}\left(\mathrm{NGenRInstalledCapTotal}_{g,s,t}\cdot\mathrm{Pgen}_{g,s}\cdot\mathrm{CV}_{g,t}\right)\geq\left(1+\mathrm{Reservemargin}\right)\cdot\mathrm{MaxLoadMW}_t\qquad\forall t$$

TotalOpreservedef$_{h,d,t}$:

$$\sum_{s,ThermalSpin_g}(\mathrm{SpinCap}_{g,s,h,d,t})+\sum_{s,ThermalQstart_g}(\mathrm{Qstartcap}_{g,s,h,d,t})\geq\mathrm{ORTot}_{h,d}\cdot\sum_s(\mathrm{Load}_{s,h,d,t})$$
$$\forall h,d,t$$

Spinningreservedef$_{h,d,t}$:

$$\sum_{s,ThermalSpin_g}(\mathrm{SpinCap}_{g,s,h,d,t})+\geq\mathrm{ORspin}_{h,d}\cdot\sum_s(\mathrm{Load}_{s,h,d,t})\qquad\forall h,d,t$$

Qstartfracdef$_{g,s,h,d,t}$:

$$\mathrm{Qstartcap}_{g,s,h,d,t}\leq(\mathrm{NGenTInstalledCapTotal}_{g,s,t}-\mathrm{NGenOn}_{g,s,h,d,t})\cdot\mathrm{Qstartfrac}_g\cdot\mathrm{Pgen}_{g,s}$$
$$\forall g,s,h,d,t\mid\mathrm{ThermalQstart}_g$$

Spinfracdef$_{g,s,h,d,t}$:

$$\mathrm{SpinCap}_{g,s,h,d,t}\leq\mathrm{NGenOn}_{g,s,h,d,t}\cdot\mathrm{Spinfrac}_g\cdot\mathrm{Pgen}_{g,s}\qquad\forall g,s,h,d,t\mid\mathrm{ThermalSpin}_g$$

OpcapUB$_{g,s,h,d,t}$:

$$\mathrm{AvgPower}_{g,s,h,d,t}+\mathrm{SpinCap}_{g,s,h,d,t}\leq\mathrm{NGenOn}_{g,s,h,d,t}\cdot\mathrm{Pgen}_{g,s}\qquad\forall g,s,h,d,t\mid\mathrm{Thermal}_g$$

OpcapLB$_{g,s,h,d,t}$:

$$\text{AvgPower}_{g,s,h,d,t} \geq \text{NGenOn}_{g,s,h,d,t} \cdot \text{Pgenmin}_g \cdot \text{Pgen}_{g,s} \qquad \forall g,s,h,d,t \mid \text{Thermalbase}_g$$

RenewoutputUB$_{g,s,h,d,t}$:

$$\text{AvgPower}_{g,s,h,d,t} = \text{NGenRInstalledCapTotal}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot \text{Cf}_{g,s,h,d} \qquad \forall g,s,h,d,t \mid \text{Renew}_g$$

NGenOnUB$_{g,s,h,d,t}$:

$$\text{NGenOn}_{g,s,h,d,t} \leq \text{NGenTInstalledCapTotal}_{g,s,t} \qquad \forall g,s,h,d,t \mid \text{Thermal}_g$$

NGenInstallTUB$_{g,t}$:

$$\sum_s (\text{NGenInstallT}_{g,s,t}) \leq \text{Nmaxgeninstall}_{g,t} \cdot (\text{Yearrel}_{t+1} - \text{Yearrel}_t)[(\text{ord(t)} < |t|)] +$$
$$\text{Nmaxgeninstall}_{g,t} \cdot (\text{Yearrel}_t - \text{Yearrel}_{t-1})[(\text{ord(t)} = |t|)] \qquad \forall g,t \mid \text{Newthermal}_g$$

NGenInstallRUB$_{g,t}$:

$$\sum_s (\text{NGenInstallR}_{g,s,t}) \leq \text{Nmaxgeninstall}_{g,t} \cdot (\text{Yearrel}_{t+1} - \text{Yearrel}_t)[(\text{ord(t)} < |t|)] +$$
$$\text{Nmaxgeninstall}_{g,t} \cdot (\text{Yearrel}_t - \text{Yearrel}_{t-1})[(\text{ord(t)} = |t|)] \qquad \forall g,t \mid \text{NewRenew}_g$$

RPSconstr$_t$:

$$\sum_{s,h,d,Renew_g} (\text{AvgPower}_{g,s,h,d,t} - \text{Curtail}_{s,h,d,t}) \cdot \text{Seasonscale}_d \geq \text{RPSfrac}_t \cdot \sum_{s,h,d} (\text{Load}_{s,h,d,t} \cdot \text{Seasonscale}_d)$$
$$\forall t \mid (\text{RPSfrac}_t > 0)$$

RPScapacityconstr$_t$:

$$\sum_{s,Renew_g} (\text{NGenRInstalledCapTotal}_{g,s,t} \cdot \text{Pgen}_{g,s}) \geq \text{RPSCapacityMin}_t$$
$$\forall t \mid (\text{RPSCapacityMin}_t > 0)$$

ObjInvgenTcostdef$_t$:

$$\text{ObjInvgenTcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \sum_{s,g,Validloc_s,Newthermal_g} (\text{NGenInstallT}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot (\text{Capmult}_g \cdot$$

$\text{Investmult}_{g,t} \cdot \text{CapCost}_{g,t} \cdot 1000 + \text{Investmult}_{g,t} \cdot \text{Gridconnect}_{g,t} \cdot 1000)) \quad \forall t \mid \text{ActiveVarSet}_t$

ObjInvgenRcostdef$_t$:

$$\text{ObjInvgenRcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \sum_{s,NewRenew_g} (\text{NGenInstallR}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot (\text{Capmult}_g \cdot \text{Investmult}_{g,t} \cdot \text{CapCost}_{g,t} \cdot 1000 + \text{Investmult}_{g,t} \cdot \text{Gridconnect}_{g,t} \cdot 1000)) \quad \forall t$$

ObjRetrofitcostdef$_t$:

$$\text{ObjRetrofitcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \Big(\sum_{s,Thermal_g} (\text{NGenTExtend}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot \text{Investmult}_{g,t} \cdot \text{LifeExtensionCost}_{g,t} \cdot 1000) + \sum_{s,Renew_g} (\text{NGenRExtend}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot \text{Investmult}_{g,t} \cdot \text{LifeExtensionCost}_{g,t} \cdot 1000)\Big) \quad \forall t$$

ObjFOMTcostdef$_t$:

$$\text{ObjFOMTcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \text{Opexmult}_t \cdot \sum_{s,Thermal_g} (\text{NGenTInstalledCapTotal}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot \text{FOMCost}_{g,t} \cdot 1000) \quad \forall t$$

ObjFOMRcostdef$_t$:

$$\text{ObjFOMRcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \text{Opexmult}_t \cdot \sum_{s,Renew_g} (\text{NGenRInstalledCapTotal}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot \text{FOMCost}_{g,t} \cdot 1000) \quad \forall t$$

ObjVOMTcostdef$_t$:

$$\text{ObjVOMTcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \text{Opexmult}_t \cdot \sum_{s,h,d,Thermal_g} (\text{AvgPower}_{g,s,h,d,t} \cdot \text{Dur}_h \cdot \text{VOMCost}_{g,t} \cdot \text{Seasonscale}_d) \quad \forall t$$

ObjFuelTcostdef$_t$:

$$\text{ObjFuelTcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \text{Opexmult}_t \cdot \sum_{s,h,d,Thermal_g} \Big(\text{AvgPower}_{g,s,h,d,t} \cdot \text{Dur}_h \cdot \frac{\text{Heatrate}_{g,s} \cdot \text{FuelCost}_{g,t}}{1000} \cdot \text{Seasonscale}_d\Big) \quad \forall t$$

ObjVOMRcostdef$_t$:

$$\text{ObjVOMRcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \text{Opexmult}_t \cdot \sum_{s,h,d,Renew_g} (\text{AvgPower}_{g,s,h,d,t} \cdot \text{Dur}_h \cdot \text{VOMCost}_{g,t} \cdot \text{Seasonscale}_d) \qquad \forall t$$

ObjEnvgencostdef$_t$:

$$\text{ObjEnvgencost}_t = \frac{\text{Opexmult}_t}{\text{Objval\_Units}} \cdot \Big(\sum_{s,h,d,} \Big(\frac{\frac{\text{AvgPower}_{g,s,h,d,t} \cdot \text{Dur}_h \cdot \text{Seasonscale}_d \cdot \text{Heatrate}_{g,s}}{1000} \cdot \text{Emissionf}_g \cdot \text{CarbonTax}_t}{1000} \Big) + \sum_{s,h,d} \Big(\frac{\text{NStartUp}_{g,s,h,d,t} \cdot \text{Pgen}_{g,s} \cdot \text{FuelCost}_{g,t} \cdot \text{StartupFueluse}_g \cdot \text{Seasonscale}_d \cdot \text{Emissionf}_g}{1000} \cdot \text{CarbonTax}_t \Big) \Big) \qquad \forall t$$

ObjStartupcostdef$_t$:

$$\text{ObjStartupcost}_t = \frac{1}{\text{Objval\_Units}} \cdot \text{Opexmult}_t \cdot \sum_{s,h,d,Thermalbase_g} \big(\text{NStartUp}_{g,s,h,d,t} \cdot \text{Pgen}_{g,s} \cdot \text{StartUpcost}_{g,t} \cdot \text{Seasonscale}_d + \text{NStartUp}_{g,s,h,d,t} \cdot \text{Pgen}_{g,s} \cdot \text{FuelCost}_{g,t} \cdot \text{StartupFueluse}_g \cdot \text{Seasonscale}_d \big) \qquad \forall t$$

Objcurtailcostdef$_t$: $\text{ObjCurtailcost}_t = \frac{\text{Opexmult}_t}{\text{Objval\_Units}} \cdot \text{Curtailcost} \cdot \sum_{s,h,d} (\text{Curtail}_{s,h,d,t} \cdot \text{Dur}_h \cdot \text{Seasonscale}_d) \qquad \forall t$

ObjCreditsdef$_t$:

$$\text{ObjCredits}_t = \frac{\text{Opexmult}_t}{\text{Objval\_Units}} \cdot \Big(\sum_{g,s,h,d,tt|(\text{Wind}_g \wedge \text{NewRenew}_g)} \big(\text{NGenInstallR}_{g,s,tt} \cdot \text{Pgen}_{g,s} \cdot \text{Cf}_{g,s,h,d} \cdot \text{Dur}_h \cdot \text{Seasonscale}_d \cdot \text{WindPTC}_{t,tt} \big) + \sum_{g,s,h,d|(\text{Wind}_g \wedge \text{ExistingRenew}_g)} \big(\text{PTCEligiblePlants}_{g,s,t} \cdot \text{Pgen}_{g,s} \cdot \text{Cf}_{g,s,h,d} \cdot \text{Dur}_h \cdot \text{Seasonscale}_d \cdot \text{PTCforExistingPlants} \big) \Big) \qquad \forall t$$

Decision variable	**Index domain**
$\text{ObjInvgenTcost}_t \geq 0$	$\forall t$
$\text{ObjInvgenRcost}_t \geq 0$	$\forall t$
$\text{ObjRetrofitcost}_t \geq 0$	$\forall t$
$\text{ObjFOMTcost}_t \geq 0$	$\forall t$
$\text{ObjFOMRcost}_t \geq 0$	$\forall t$
$\text{ObjVOMTcost}_t \geq 0$	$\forall t$
$\text{ObjFuelTcost}_t \geq 0$	$\forall t$
$\text{ObjFuelRcost}_t \geq 0$	$\forall t$
$\text{ObjVOMRcost}_t \geq 0$	$\forall t$
$\text{ObjEnvgencost}_t \geq 0$	$\forall t$
$\text{ObjStartupcost}_t \geq 0$	$\forall t$
$\text{ObjCurtailcost}_t \geq 0$	$\forall t$
$\text{ObjCredits}_t \geq 0$	$\forall t$
$\text{AvgPower}_{g,s,h,d,t} \geq 0$	$\forall g, s, h, d, t$
$\text{Curtail}_{s,h,d,t} \geq 0$	$\forall s, h, d, t$
$\text{NGenOn}_{g,s,h,d,t} \in \mathbb{Z}_+$	$\forall g, s, h, d, t$
$\text{NStartUp}_{g,s,h,d,t} \in \mathbb{Z}_+$	$\forall g, s, h, d, t$
$\text{NShutD}_{g,s,h,d,t} \in \mathbb{Z}_+$	$\forall g, s, h, d, t$
$\text{NGenT}_{g,s,t} \in \mathbb{Z}_+$	$\forall g, s, t$
$\text{NGenRetireT}_{g,s,t} \in \mathbb{Z}_+$	$\forall g, s, t$
$\text{NGenTExtend}_{g,s,t} \in \mathbb{Z}_+$	$\forall g, s, t$
$\text{NGenTInstalledCapTotal}_{g,s,t} \in \mathbb{Z}_+$	$\forall g, s, t$
$\text{NGenInstallT}_{g,s,t} \in \mathbb{Z}_+$	$\forall g, s, t$
$\text{NGenR}_{g,s,t} \geq 0$	$\forall g, s, t$
$\text{NGenRetireR}_{g,s,t} \geq 0$	$\forall g, s, t$
$\text{NGenRExtend}_{g,s,t} \geq 0$	$\forall g, s, t$
$\text{NGenRInstalledCapTotal}_{g,s,t} \geq 0$	$\forall g, s, t$
$\text{NGenInstallR}_{g,s,t} \geq 0$	$\forall g, s, t$
$\text{Qstartcap}_{g,s,h,d,t} \geq 0$	$\forall g, s, h, d, t$

A.1.2 Time Slice Capacity Expansion Model (TS-GEP)

Notes: The time slice model includes a set of nodes, for a transmission network, and a set of states. Neither are used the study. That is, each set is a singleton.

Sets

Name	Domains	Description
block, block1, block2	*	Set of blocks associated with time load blocks
g	g	Set of generator types and technologies (existing and potential)
n	*	Set of nodes
t, tt	t	Set of all time periods
state	*	Set of states
fuel_type	*	Set of fuel types
fuel_bin	*	Set of fuel bins for the fuel supply curves
supply_bin	*	Set of supply bins for each resource capacity supply curves
t_ptc	t	Set of time periods (years) in which the production tax credit
g_thermal	g	Set of all (old and new) thermal generators (e.g., Coal, NG, Nuclear)
g_renew	g	Set of renewable generators
g_dispatchable	g	Set of dispatchable generators that can adjust power output on demand and contribute to operating reserves
g_wind	g_renew	Set of wind generators
g_wind_new	g_wind	Set of new wind generators
g_new	g	Set of new generator types
g_itc_eligible	g_renew	Set of ITC eligible generator types

Parameters

Name	Domains	Description
CapacityFactor	g, n, t, block	The fraction of nameplate capacity that is actually available for generator type g at node n in time period t in time block 'block'
CapacityValue	g	Amount of electrical demand that may be added in each time-slice for an incremental increase in capacity of a given VRRE technology
CapitalMultiplier	g	Capital multiplier associated with generator type g (positive scalar typically between 1 and 2
ExistingInstalledCapacity	g, n	Existing installed capacity of generator type g at node n prior to the planning horizon (MW)
GHG	g	GHG emissions of generator (ton CO2e per MMBtu)
HeatRate	g	Heat rate of generator type g (MMBtu per MWh)
LeadTime	g	Lead time between decision to invest and the time when generator type g is operational (years)
LifeTime	g	Lifetime of generator type g (years)
MTDF	g	Minimum turndown fraction (as a fraction of nameplate capacity) that a generator must maintain (e.g. for nuclear generators ReEDS assumes this parameter is 1.00 implying that nuclear generators must run at capacity when they are available)

Name	Domains	Description
ExtensionCost	g, t	Cost to extend indefinitely the lifetime of generator type g in time period t ($ per MW)
FOMCost	g, t	Annual fixed operating & maintenance costs for generator type g in time period t ($ per MWh)
GridConnectCost	g, t	Cost of connecting new generation capacity of generator type g in time period t to the grid ($ per MW)
Min_Cumulative_MW_To_Extend_Or_Retire	g, t	Minimum capacity (MW) of generator type g to extend or retire by time period t
VOMCost	g, t	Variable operating & maintenance costs for generator type g in time period t ($ per MWh)
FuelTypeCost_Bin	fuel_type, t, fuel_bin	Cost of fuel type fuel_type in time period t in fuel bin fuel_bin at ($ per MMBtu)
GenInstallUB	g, n, t	Upper bound on amount of generation type g that can be installed at node n in time period t (MW)
GenInstallCost_Bin	g, t, supply_bin	Overnight installation capital cost of generator type g in time period t in supply_bin ($ per MW)
LoadMW	n, t, block	Average load at node n in time period t in load block 'block' (MW)
LoadMWh	n, t, block	Average load at node n in time period t in load block 'block' (MWh)
MaxLoadMW	n, t	Maximum load at node n in time period t (MW)
NumHours	block	Number of hours in time period t in load block block (positive integer)

Name	Domains	Description
PlanningReserveMargin	n, t	Fraction (e.g., 0.1375) of max load that must be met in the capacity reserve constraints
OperatingReserveMargin		Fraction (e.g., 0.075) of load that must be met by spinning and quickstart reserves
MaxQuickstartReserveMargin		Fraction (e.g., 0.06) denoting the maximum amount of operating reserves that can be supplied from quickstarts
MinSpinningReserveMargin		Fraction (e.g., 0.03) denoting the minimum amount of amount of spinning reserves
ForecastErrorReserve		Fraction
RPS_MIN_GEN_AMOUNT	state, t	Minimum generation (MWh) in state 'state' in time period t required by Renewable Portfolio Standards
BlocksBelongToSameSeason	block1, block2	1 if blocks belong to same season - 0 otherwise
IsFuelTypeGeneratorPair	fuel_type, g	1 if generator type g uses fuel type fuel_type - 0 otherwise (Needed for supply curves)
IsStateNodePair	state, n	1 if node n is in state - 0 otherwise
CarbonPrice	t	Carbon price ($ per ton C02e)
ITC_Fraction	g, t	Investment tax credit fraction (fraction of capital cost) applied to generator type g installed in time period t
Opexmult	t	Operating multiplier in time period t to account for intermediate model years
PTC	t_ptc	Production tax credit ($ per MWh) to apply for new wind generators installed in time period t_ptc

Name	Domains	Description
rel_ord	t	Relative order of time period t e.g. rel_ord('t3') = 3 even if ord('t3') = 2
PTC_constant	t	Constant ($) of production tax credit attributed to wind-old in time period t (computed directly in GAMS)
DiscountFactor		Scalar (Scalar between 0 and 1)
Include_Tax_Credits		Takes value 1 if investment and production tax credits should be included - 0 otherwise
PenaltyUnmetDemand		Penalty for each unit of unmet demand ($ per MWh)
PenaltyExcessGeneration		Penalty for each unit of excess generation ($ per MWh)
Objval_Units		Units in which the objective function value is measured (e.g. 1e9 ==>US$1 billion)

Variables

Name	Domains	Description
avgPower	g, n, t, block	Average power (capacity in use) from generator g at node n in time period t in load block block (MW)
generation	g, n, t, block	Average generation from g from generator g at node n in time period t in load block block (MWh)
genInstall	g, n, t	Capacity (MW) of generation type g initially installed at node n in time period t
genInstall_Bin	g, t, supply_bin	Capacity (MW) of generation type g installed in time period t in supply bin 'supply_bin'

Name	Domains	Description
installedCapacity	g, n, t	Total installed nameplate capacity (MW) of generator type g at node n available to serve load in time period t (after all installation - upgrades - and retirements take place in time period t)
effectiveCapacity	g, n, t, block	Effective or firm capacity (MW) of generator type g at node n available to serve load in time period t
extendedCapacity	g, n, t	Capacity (MW) of generation type g at node n in time period t whose lifetime is extended indefinitely
retiredCapacity	g, n, t	Capacity (MW) of generation type g retired at node n in time period t
spinningReserve	g_dispatchable, n, t, block	Power (MW) of dispatchable generator g available to serve spinning reserves at node n in time period t in block block (MW)
quickstartReserve	g_dispatchable, n, t, block	Power (MW) of dispatchable generator g available for quickstart at node n in time period t in block block (MW)
amountEnergyConsumed	g, t	Amount of energy (MMBtu) consumed by generator type g in time period t
amountFuelTypeConsumed	fuel_type, t	Amount of fuel type fuel_type consumed in time period t (MMBtu)
amountFuelTypeConsumed InFuelBin	fuel_type, t, fuel_bin	Amount of fuel type fuel_type consumed in time period t in fuel bin fuel_bin (Quad = 1e9 MMBtu = 1e15 Btu)
unmetDemand	n, t, block	Unmet demand (MWh)
excessGeneration	n, t, block	Excess generation (MWh)

Name	Domains	Description
ptc_generation	g_wind_new, t_ptc, t	Generation (MWh) from new wind resources in time period t that were built in time period t_ptc
objval		Objective function value
VC_Obj_InstallationCost	t	Auxiliary variable to isolate installation costs
VC_Obj_AnnualFOMCost	t	Auxiliary variable to isolate annual FOM costs
VC_Obj_CarbonTax	t	Auxiliary variable to isolate emissions costs
VC_Obj_ExtensionCost	t	Auxiliary variable to isolate extension costs
VC_Obj_FuelCost	t	Auxiliary variable to isolate fuel costs
VC_Obj_PenaltyUnmetDemand	t	Auxiliary variable to isolate unmet demand costs
VC_Obj_PenaltyExcessGen	t	Auxiliary variable to isolate excess generation costs
VC_Obj_VOMCost	t	Auxiliary variable to isolate VOM costs
VC_Obj_ITC	t	Auxiliary variable to isolate investment tax credit savings
VC_Obj_PTC	t	Auxiliary variable to isolate production tax credit savings

Equations

Name	Domains	Description
C_ObjectiveFunction		Objective function minimizing generation cost
C_LoadConstr	n, t, block	Load constraint at node n in time period t
C_PowerToGeneration Constr	g, n, t, block	Convert average power (MW) to generation (MWh) in block
C_InstallationCostConstr	t	Redundant constraint to isolate installation costs

Name	Domains	Description
C_AnnualFixedOMCost Constr	t	Redundant constraint to isolate annual FOM costs
C_VarOMCostConstr	t	Redundant constraint to isolate VOM costs
C_FuelCostConstr	t	Redundant constraint to isolate fuel costs
C_ExtensionCostConstr	t	Redundant constraint to isolate extension costs
C_CarbonTaxConstr	t	Redundant constraint to isolate emissions costs
C_PenaltyUnmetDemand Constr	t	Redundant constraint to isolate unmet demand costs
C_PenaltyExcessGeneration Constr	t	Redundant constraint to isolate excess generation costs
C_ITC_Obj_Constr	t	Redundant constraint to isolate investment tax credit savings
C_PTC_Obj_Constr	t	Redundant constraint to isolate production tax credit savings
C_InstalledCapConstr	g, n, t	Capacity balance constraint for each generator type g in each time period t
C_EffectiveCapConstr	g, n, t, block	The effective capacity of generator type g in time period t and time block 'block' equals the capacity factor in that block times the installed capacity
C_RetireCapConstr	g, n, t	Retire or extend capacity that exceeds its lifetime + leadtime to build
C_DispatchableGenCap Constr	g_dispatchable, n, t, block	Dispatchable power plus spinning reserves plus quickstart reserves must not exceed effective capacity
C_RenewableGenCap Constr	g_renew, n, t, block	Average power equals effective capacity

Name	Domains	Description
C_MinimumTurndown Constr	g_thermal, n, t, block1, block2	Minimum power level that a generator must satisfy in each time block
C_PlanningReserveConstr	n, t	Planning reserve requirements at node n in time period t
C_OperatingReserve Constr	n, t, block	Operating reserve requirement at node n in time period t in load block block - ReEDS does this for each rs_group
C_MaxQuickstartReserve Constr	n, t, block	
C_SpinningReserveConstr	n, t, block	Spinning reserve requirement at node n in time period t in load block block - ReEDS does this for each rs_group
C_RPS_Gen_Constr	state, t	Requires a minimum generation amount (MWh) from renewable resources in a particular state in time period t
C_AmountEnergyConsumed Constr	g, t	The amount of energy consumed by generator type g in time period t equals heat rate times the generation of g over all time blocks in that period
C_FuelTypeConsumed Constr	fuel_type, t	Equate amount of fuel type fuel_type consumed in time period t with the amount of generation from that same fuel type
C_FuelTypeConsumed SupplyCurveConstr	fuel_type, t	Equate amount of fuel type fuel_type consumed in time period t with amount of fuel type fuel_type consumed in each fuel bin

Name	Domains	Description
C_InstallationSupply CurveConstr	g, t	Sum of installations over all supply bins must equal the total amount installed (for each g,t pair)
C_PTC_Constr	g_wind_new, t_ptc, t	Computes the correct amount of ptc eligible generation

Equation Definitions

C_ObjectiveFunction:

$$\sum_t (\text{DiscountFactor}^{\text{rel\_ord}_t} \cdot (\text{VC\_Obj\_InstallationCost}_t + \text{VC\_Obj\_AnnualFOMCost}_t + \text{VC\_Obj\_VOMCost}_t + \text{VC\_Obj\_FuelCost}_t + \text{VC\_Obj\_CarbonTax}_t + \text{VC\_Obj\_PenaltyUnmetDemand}_t + \text{VC\_Obj\_PenaltyExcessGen}_t + \text{VC\_Obj\_ExtensionCost}_t - \text{VC\_Obj\_ITC}_t[\text{Include\_Tax\_Credits}] - \text{VC\_Obj\_PTC}_t[\text{Include\_Tax\_Credits}])) = \text{objval}$$

C_InstallationCostConstr$_t$: c

$$\text{VC\_Obj\_InstallationCost}_t = \sum_{g,supply\_bin} ((\text{CapitalMultiplier}_g \cdot \text{GenInstallCost\_Bin}_{g,t,supply\_bin} + \text{GridConnectCost}_{g,t}) \cdot \text{genInstall\_Bin}_{g,t,supply\_bin}) \cdot \frac{1}{\text{Objval\_Units}} \qquad \forall t$$

C_AnnualFixedOMCostConstr$_t$:

$$\text{VC\_Obj\_AnnualFOMCost}_t = \text{Opexmult}_t \cdot \sum_{g,n} (\text{FOMCost}_{g,t} \cdot \text{installedCapacity}_{g,n,t}) \cdot \frac{1}{\text{Objval\_Units}} \qquad \forall t$$

C_VarOMCostConstr$_t$:

$$\text{VC\_Obj\_VOMCost}_t = \text{Opexmult}_t \cdot \sum_{g,n,block} (\text{VOMCost}_{g,t} \cdot \text{generation}_{g,n,t,block}) \cdot \frac{1}{\text{Objval\_Units}} \qquad \forall t$$

C_ExtensionCostConstr$_t$:

$$\text{VC\_Obj\_ExtensionCost}_t = \sum_{g,n} (\text{ExtensionCost}_{g,t} \cdot \text{extendedCapacity}_{g,n,t}) \cdot \frac{1}{\text{Objval\_Units}} \qquad \forall t$$

C_FuelCostConstr$_t$:

$$\text{VC\_Obj\_FuelCost}_t = \frac{1}{\text{Objval\_Units}} \cdot \text{Opexmult}_t \cdot \sum_{fuel\_type,fuel\_bin} (\text{FuelTypeCost\_Bin}_{fuel\_type,t,fuel\_bin} \cdot \text{amountFuelTypeConsumedInFuelBin}_{fuel\_type,t,fuel\_bin}) \quad \forall t$$

C_CarbonTaxConstr$_t$:

$$\text{VC\_Obj\_CarbonTax}_t = \text{Opexmult}_t \cdot \sum_{g} \big(\text{CarbonPrice}_t \cdot \text{GHG}_g \cdot \text{amountEnergyConsumed}_{g,t}\big) \cdot \frac{1}{\text{Objval\_Units}} \quad \forall t$$

C_PenaltyUnmetDemandConstr$_t$:

$$\text{VC\_Obj\_PenaltyUnmetDemand}_t = \text{Opexmult}_t \cdot \sum_{n,block} \big(\text{PenaltyUnmetDemand} \cdot \text{unmetDemand}_{n,t,block}\big) \cdot \frac{1}{\text{Objval\_Units}} \quad \forall t$$

C_PenaltyExcessGenerationConstr$_t$:

$$\text{VC\_Obj\_PenaltyExcessGen}_t = \text{Opexmult}_t \cdot \sum_{n,block} \big(\text{PenaltyExcessGeneration} \cdot \text{excessGeneration}_{n,t,block}\big) \cdot \frac{1}{\text{Objval\_Units}} \quad \forall t$$

C_ITC_Obj_Constr$_t$:

$$\text{VC\_Obj\_ITC}_t = \sum_{g\_itc\_eligible,supply\_bin} (\text{ITC\_Fraction}_{g\_itc\_eligible,t} \cdot \text{CapitalMultiplier}_{g\_itc\_eligible} \cdot \text{GenInstallCost\_Bin}_{g\_itc\_eligible,t,supply\_bin} \cdot \text{genInstall\_Bin}_{g\_itc\_eligible,t,supply\_bin}) \cdot \frac{1}{\text{Objval\_Units}} \quad \forall t \mid \text{Include\_Tax\_Credits}$$

C_PTC_Obj_Constr$_t$:

$$\text{VC\_Obj\_PTC}_t = \text{Opexmult}_t \cdot \Big(\text{PTC}_{\texttt{t1}} \cdot \text{PTC\_constant}_t + \sum_{g\_wind\_new,t\_ptc|(\text{ord(t\_ptc)} \leq \text{rel\_ord}_t)} \big(\text{PTC}_{t\_ptc} \cdot \text{ptc\_generation}_{g\_wind\_new,t\_ptc,t}\big)\Big) \cdot \frac{1}{\text{Objval\_Units}} \quad \forall t \mid \text{Include\_Tax\_Credits}$$

C_LoadConstr$_{n,t,block}$**:**

$$\sum_{g}(\text{generation}_{g,n,t,block}) + \text{unmetDemand}_{n,t,block} = \text{LoadMWh}_{n,t,block} + \text{excessGeneration}_{n,t,block}$$

$\forall n, t, block$

C_PowerToGenerationConstr$_{g,n,t,block}$**:**

$$\text{generation}_{g,n,t,block} = \text{avgPower}_{g,n,t,block} \cdot \text{NumHours}_{block} \qquad \forall g, n, t, block$$

C_EffectiveCapConstr$_{g,n,t,block}$**:**

$$\text{effectiveCapacity}_{g,n,t,block} = \text{CapacityFactor}_{g,n,t,block} \cdot \text{installedCapacity}_{g,n,t} \qquad \forall g, n, t, block$$

C_InstalledCapConstr$_{g,n,t}$**:**

$$\text{installedCapacity}_{g,n,t} = \text{ExistingInstalledCapacity}_{g,n}[(\text{rel\_ord}_t = 1)] + \text{installedCapacity}_{g,n,t-1} + \text{genInstall}_{g,n,t-\text{LeadTime}_g} - \text{retiredCapacity}_{g,n,t} \qquad \forall g, n, t$$

C_RetireCapConstr$_{g,n,t}$**:**

$$\sum_{tt|(\text{rel\_ord}_{tt} \leq \text{rel\_ord}_t)} (\text{extendedCapacity}_{g,n,tt} + \text{retiredCapacity}_{g,n,tt})$$

$$\geq \text{Min\_Cumulative\_MW\_To\_Extend\_Or\_Retire}_{g,t}[(\text{ExistingInstalledCapacity}_{g,n} > 0)]$$

$$+ \sum_{tt|(\text{rel\_ord}_{tt} \leq (\text{rel\_ord}_t - (\text{LeadTime}_g + \text{LifeTime}_g)))} (\text{genInstall}_{g,n,tt})$$

$\forall g, n, t \mid ((\text{ExistingInstalledCapacity}_{g,n} > 0) \vee \text{g\_new}_g \wedge ((\text{rel\_ord}_t - (\text{LeadTime}_g + \text{LifeTime}_g)) > 0))$

C_DispatchableGenCapConstr$_{g\_dispatchable,n,t,block}$**:**

$$\text{avgPower}_{g\_dispatchable,n,t,block} + \text{spinningReserve}_{g\_dispatchable,n,t,block} + \text{quickstartReserve}_{g\_dispatchable,n,t,block} \leq \text{effectiveCapacity}_{g\_dispatchable,n,t,block}$$

$\forall g\_dispatchable, n, t, block$

C_RenewableGenCapConstr$_{g\_renew,n,t,block}$**:**

$$\text{avgPower}_{g\_renew,n,t,block} = \text{effectiveCapacity}_{g\_renew,n,t,block} \qquad \forall g\_renew, n, t, block$$

C_MinimumTurndownConstr$_{g\_thermal,n,t,block1,block2}$**:**

$$\text{avgPower}_{g\_thermal,n,t,block1} \geq \text{MTDF}_{g\_thermal} \cdot \text{avgPower}_{g\_thermal,n,t,block2}$$
$$\forall g\_thermal, n, t, block1, block2 \mid ((\text{MTDF}_{g\_thermal} > 0) \wedge (\neg(\text{ord}(\text{block1}) = \text{ord}(\text{block2}))) \wedge \text{BlocksBelongToSameSeason}_{block1,block2})$$

C_PlanningReserveConstr$_{n,t}$**:**

$$\sum_{g} (\text{CapacityValue}_g \cdot \text{installedCapacity}_{g,n,t}) \geq \text{MaxLoadMW}_{n,t} \cdot (1 + \text{PlanningReserveMargin}_{n,t}) \quad \forall n, t$$

C_OperatingReserveConstr$_{n,t,block}$**:**

$$\sum_{g\_dispatchable} (\text{spinningReserve}_{g\_dispatchable,n,t,block} + \text{quickstartReserve}_{g\_dispatchable,n,t,block})$$
$$\geq \text{LoadMW}_{n,t,block} \cdot \text{OperatingReserveMargin} + \text{ForecastErrorReserve} \qquad \forall n, t, block$$

C_MaxQuickstartReserveConstr$_{n,t,block}$**:**

$$\sum_{g\_dispatchable} (\text{quickstartReserve}_{g\_dispatchable,n,t,block}) \leq \text{LoadMW}_{n,t,block} \cdot \text{MaxQuickstartReserveMargin} + \frac{5}{6} \cdot \text{ForecastErrorReserve} \qquad \forall n, t, block$$

C_SpinningReserveConstr$_{n,t,block}$**:**

$$\sum_{g\_dispatchable} (\text{spinningReserve}_{g\_dispatchable,n,t,block})$$
$$\geq \text{LoadMW}_{n,t,block} \cdot \text{MinSpinningReserveMargin} \qquad \forall n, t, block$$

C_RPS_Gen_Constr$_{state,t}$**:**

$$\sum_{g\_renew,n,block \mid \text{IsStateNodePair}_{state,n}} (\text{generation}_{g\_renew,n,t,block})$$
$$- \sum_{n,block \mid \text{IsStateNodePair}_{state,n}} (\text{excessGeneration}_{n,t,block})$$
$$\geq \text{RPS\_MIN\_GEN\_AMOUNT}_{state,t} \qquad \forall state, t \mid (\text{RPS\_MIN\_GEN\_AMOUNT}_{state,t} > 0)$$

C_AmountEnergyConsumedConstr$_{g,t}$**:**

$$\text{amountEnergyConsumed}_{g,t} = \sum_{n,block} \left(\text{HeatRate}_g \cdot \text{generation}_{g,n,t,block}\right) \qquad \forall g, t$$

C_FuelTypeConsumedConstr$_{fuel\_type,t}$**:**

$$\text{amountFuelTypeConsumed}_{fuel\_type,t} = \sum_{g|\text{IsFuelTypeGeneratorPair}_{fuel\_type,g}} (\text{amountEnergyConsumed}_{g,t}) \qquad \forall fuel\_type, t$$

C_FuelTypeConsumedSupplyCurveConstr$_{fuel\_type,t}$**:**

$$\text{amountFuelTypeConsumed}_{fuel\_type,t} = \sum_{fuel\_bin} \left(\text{amountFuelTypeConsumedInFuelBin}_{fuel\_type,t,fuel\_bin}\right) \qquad \forall fuel\_type, t$$

C_InstallationSupplyCurveConstr$_{g,t}$**:**

$$\sum_{supply\_bin} \left(\text{genInstall\_Bin}_{g,t,supply\_bin}\right) = \sum_{n} (\text{genInstall}_{g,n,t}) \qquad \forall g, t$$

C_PTC_Constr$_{g\_wind\_new,t\_ptc,t}$**:**

$$\text{ptc\_generation}_{g\_wind\_new,t\_ptc,t} = \sum_{n,block} \left(\text{CapacityFactor}_{g\_wind\_new,n,t,block} \cdot \text{genInstall}_{g\_wind\_new,n,t\_ptc}\right)$$
$$\forall g\_wind\_new, t\_ptc, t \mid \left(\text{Include\_Tax\_Credits} \wedge \left(\text{rel\_ord}_{t\_ptc} \leq \text{rel\_ord}_{t}\right) \wedge \left(\text{rel\_ord}_{t} \leq (\text{rel\_ord}_{t\_ptc} + 10 - 1)\right)\right)$$

C_Installation_Limit$_{g,n,t}$**:**

$$\text{genInstall}_{g,n,t} \leq \text{GenInstallUB}_{g,n,t} \qquad \forall g, n, t$$

Decision variable	**Index domain**
$\text{genInstall\_Bin}_{g,t,supply\_bin} \geq 0$	$\forall g, t, supply\_bin$
$\text{installedCapacity}_{g,n,t} \geq 0$	$\forall g, n, t$
$\text{generation}_{g,n,t,block} \geq 0$	$\forall g, n, t, block$
$\text{extendedCapacity}_{g,n,t} \geq 0$	$\forall g, n, t$
$\text{amountFuelTypeConsumedInFuelBin}_{fuel\_type,t,fuel\_bin} \geq 0$	$\forall fuel\_type, t, fuel\_bin$
$\text{amountEnergyConsumed}_{g,t} \geq 0$	$\forall g, t$
$\text{unmetDemand}_{n,t,block} \geq 0$	$\forall n, t, block$
$\text{excessGeneration}_{n,t,block} \geq 0$	$\forall n, t, block$
$\text{ptc\_generation}_{g\_wind\_new,t\_ptc,t} \geq 0$	$\forall g\_wind\_new, t\_ptc, t$
$\text{avgPower}_{g,n,t,block} \geq 0$	$\forall g, n, t, block$
$\text{effectiveCapacity}_{g,n,t,block} \geq 0$	$\forall g, n, t, block$
$\text{genInstall}_{g,n,t} \geq 0$	$\forall g, n, t$
$\text{retiredCapacity}_{g,n,t} \geq 0$	$\forall g, n, t$
$\text{spinningReserve}_{g\_dispatchable,n,t,block} \geq 0$	$\forall g\_dispatchable, n, t, block$
$\text{quickstartReserve}_{g\_dispatchable,n,t,block} \geq 0$	$\forall g\_dispatchable, n, t, block$
$\text{amountFuelTypeConsumed}_{fuel\_type,t} \geq 0$	$\forall fuel\_type, t$

A.2 Data assumptions

Table A.10 and Table A.11 summarize technology and cost assumptions used to model operations of existing and new generators, respectively, in the two GEP models and the PCS model. Unless otherwise stated, all cost parameters reported below are reported in 2015 dollars. Some key points to note regarding the data in Table A.10 and Table A.11:

- Both GEPs do not explicitly consider the construction time for power plants. Instead the construction time is implicitly considered by accounting for the cost of capital financing during the construction period in the capital cost assumptions of each technology, following the methodology presented in the 2016 NREL technology baseline (NREL, 2016). In this manner, both GEP models implicitly distinguish between the relative construction times of different technologies. The capital multiplier associated with new generator clusters is meant to account for differences in depreciation schedules applicable to each technology in the U.S. context, with higher values being indicative of a slower depreciating schedule and vice versa.

- In the absence of better data sources, we assume the startup costs for nuclear power plants to be the same as the values reported for coal power plants.
- The capacity contribution of wind plants to the planning reserve margin is based on the average value of their capacity contribution to peak demand in summer and winter months between 2009-2014 (ERCOT, 2016b). Currently, ERCOT estimates the capacity value of solar PV plants to be 0.8-1, due to the small amount of total installed capacity (ERCOT, 2016b). We use a lower value of 0.6 to reflect the declining contribution of solar PV to peak demand with increasing installed PV capacity.
- The life extension costs for existing generators is based on a review of FERC form 1 data regarding the reported annual capital expenditures made by older units and is reported in the IPM documentation as a proportion of the capital costs of the corresponding new generators (EPA, 2013). For example, the life extension cost of an existing NGCC plant in a given year is assumed to be 9.3% of the capital cost of a new NGCC plant in that year. We assume that the life extension costs for natural gas boiler plants (NGST) to be the same as the extension costs for existing coal plants, due to the similar equipment in use (e.g. boilers, steam turbines). In all cases, the life extension costs are assumed to double the lifespan of the generator (EPA, 2013).

Table A.12 summarizes the annual installation limits assumed for wind, solar PV and solar CSP generators. These values were obtained from scaling down the annual installation limits assumed in the Integrated Planning Model (EPA, 2015) for the entire U.S., based on the relative of share of annual power generation in ERCOT. Additionally, the installation limits for the period beyond 2018 shown in Table A.12 are scaled up in the GEP models by a factor 3 to account for the fact that both the GEP models step forward in three year time increments. Under some scenarios investigated here, such as 50-70% RE scenarios,

Table A.10: Technology and cost assumptions for existing generator fleet. HHV = Higher heating value

	Source	Coal	NGCT	NGCC	NGST	Nuclear	Solar PV	Wind
Nameplate capacity (MW)	Estimated based on generator categories in (ERCOT, 2016c)	644	156	719	622	1291	39	134
Heat rate - HHV (Btu/kWh)		10484	11395	8409	13216	10479	-	-
Ramp rate up and down (% of nameplate capacity/hr)	(Kerl et al., 2015), (NWPCC, 2010), (WEC, 1984)	25%	100%	100%	25%	17%	-	-
Min. output (% of nameplate capacity)	(ERCOT, 2016c)	48%	25%	32%	28%	90%	-	-
Max. spin reserves (% of nameplate capacity)	(Mai et al., 2013)	10%	50%	10%	10%	0%	-	-
Max. quick start reserves (% of nameplate capacity)	Assumption	0%	100%	100%	0%	0%	-	-
Lifetime (years)	(Tidball et al., 2010)	60	30	30	60	60	30	20
Start-up costs ($/MW)	(Mai et al., 2013)	140.94	37.15	86.31	140.94	140.94	-	-
Start-up fuel usage (MMBtu/MW)	(Mai et al., 2013)	14.5	1.53	0.24	14.5	14.5	-	-
Fixed O&M cost ($/kW)	(ERCOT, 2016c) (NWPCC, 2010)	26.87	5.27	13.96	16.59	76.91	42.48	30.82
Variable O&M cost ($/MWh)	(ERCOT, 2016c)	5.27	4.21	3.16	6.85	4.21	-	-
Life extension cost as a proportion of new unit capital cost (%)	(EPA, 2013)	7.0	4.2	9.3	7.0	9.0	4.2	4.2
Capacity contribution to reserve margin	(ERCOT, 2016b) (Webster et al., 2013)	1	1	1	1	1	0.6	0.15
Minimum turndown fraction (only TS-GEP)	(Short et al., 2011)	0.4	0	0	0	1	-	-

the assumed installation limits for each model year are found to be binding in a few years and limit the rate of deployment of wind or solar PV technologies. The installation limits

Table A.11: Technology and cost assumptions for new generator clusters. HHV = Higher heating value

	Source	Nuclear	Wind	Solar PV	Solar CSP	Coal IGCC	Coal IGCC CCS	NGCC	NGCC CCS	NGCT
Nameplate capacity (MW)	(EIA, 2015)	2234	100	20	100	600	520	400	340	210
Heat rate - HHV (Btu/kWh)	(EIA, 2015) (GE Power, 2016)	10479				7450	8307	6260	7493	8550
Ramp rate up and down (% of nameplate capacity/hr)	(Kerl et al., 2015), (NWPCC, 2010), (WEC, 1984)	17%				25%	25%	100%	100%	100%
Min. output (% of nameplate capacity)	(GE Power, 2016), (Black & Veatch, 2012), (Truby, 2014)	50%	-	-	-	30%	30%	40%	40%	30%
Max. spin reserves (% of nameplate capacity)	(Mai et al., 2013)	-	-	-	-	10%	10%	10%	10%	50%
Max. quick start reserves (% of nameplate capacity)	Assumption	-	-	-	-	0%	0%	100%	100%	100%
Lifetime (years)	(Tidball et al., 2010)	60	20	30	30	60	60	30	30	30
Capital cost multiplier	(NREL, 2016)	1.28	1.13	1.13	1.13	1.33	1.33	1.28	1.28	1.28
Start-up costs ($/MW)	(Mai et al., 2013)	140.94	-	-	-	140.9	140.9	86.31	86.31	37.15
Start-up fuel usage (MMBtu/MW)	(Mai et al., 2013)	-	-	-	-	14.5	14.5	0.24	0.24	1.5
Fixed O&M cost ($/kW)[1]	(NREL, 2016)	94.68	46-50	8-16	51-68	52.1	73.9	14.48	32.27	7.3
Variable O&M cost ($/MWh)	(NREL, 2016)	2.17	-	-	3	7.3	8.6	3.50	6.8	13.1
Capacity contribution to reserve margin	(ERCOT, 2016b) (Webster et al., 2013)	1	0.15	0.6	0.6	1	1	1	1	1
Minimum turndown fraction (only TS-GEP)	(Short et al., 2011)	1	-	-	-	0.5	0.5	0	0	0

assumed for the remaining new generator clusters were always much larger than the installed capacity for all the scenarios considered here and therefore are not shown in Table A.12.

Figure A.1 shows the assumed capital cost projections over time for the new generator technologies considered in both GEP models. The data was derived from 2016 NREL technology baseline (NREL, 2016). As a sensitivity analysis, we considered an alternative trajectory for capital costs of solar PV over time that does not go below $1300/kW, as shown in Figure A.2.

Figure A.3 plots the fuel price projections over time considered in both GEP models

Table A.12: Annual installation limits for RE technologies. Data source: (EPA, 2015)

	2015-2018	2018-2021	2021-2024	2024-2027	2027-2030	2030-2033	2033-2036	2036-2039	2039-2042	2042-2045
Wind	1570	3139.9	3139.9	3139.9	7849.8	7849.8	7849.8	7849.8	7849.8	7849.8
Solar PV	744.1	1488.2	1488.2	1488.2	3720.6	3720.6	3720.6	3720.6	3720.6	3720.6
Solar CSP	900	900	900	900	900	900	900	900	900	900

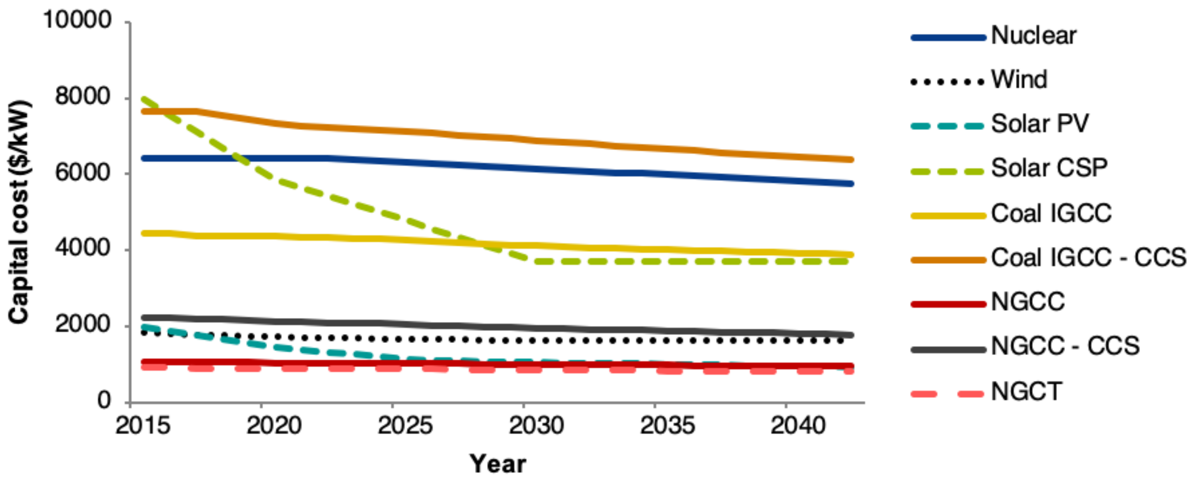

Figure A.1: CAPEX over time for all new generator types (source: NREL (2016)).

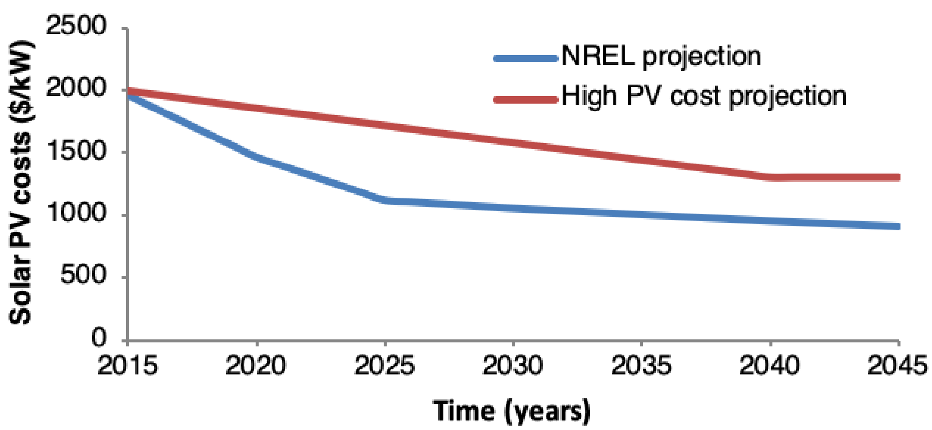

Figure A.2: High solar PV cost projections (input to high solar PV cost scenario set shown in Figure 4.8C) relative the NREL projections (NREL, 2016) used in the reference scenario set.

and the PCS model. The data was derived from EIA Annual Energy Outlook 2016 (EIA, 2016). Unless otherwise stated, all results presented consider the "reference" set of fuel price projections.

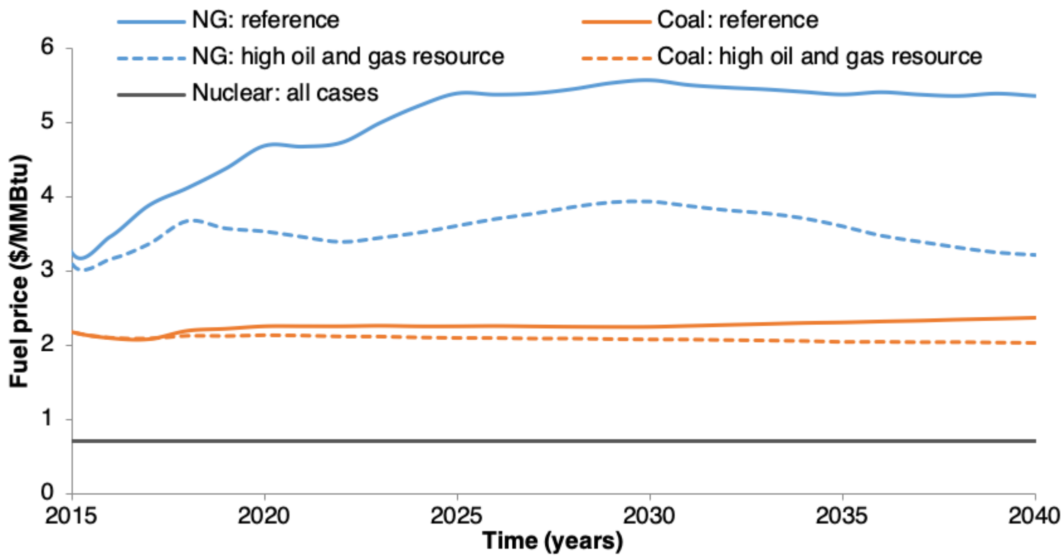

Figure A.3: Fuel price projections for "reference" scenarios and "high oil and gas resource" scenarios. Data source: EIA Annual Energy Outlook 2016 (EIA, 2016). All values reported on a Higher Heating Value (HHV) basis.

Table A.13 summarizes the implementation of the investment tax credits for wind and solar PV technologies in the two GEP models that approximates the current policy (DOE, 2016a). In each case, the investment tax credit effectively reduces the capital cost by the specified percentage. The production tax credit (PTC) is also incorporated in both GEP models according to current policy (DOE, 2016b). Specifically, the PTC is available for wind generators, both existing and new, constructed before 2019. For each plant, the PTC is available for the first 10 years of their operation. Additionally, the PTC of plants built in the 2018-2021 period is 60% of the current PTC value, i.e. $23/MWh. It should be noted that we did not consider the PTC for the PCS model runs for different RE scenarios.

Table A.13: Investment tax credits for new installations of wind and solar PV technologies as % percentage of capital cost implemented in the two GEP models. Data source: DOE (2016a)

	2015-2018	2018-2021	2021-2024	2024-2027	2027-2030	2030-2033	2033-2036	2036-2039	2039-2042	2042-2045
Wind	30%	18%	0%	0%	0%	0%	0%	0%	0%	0%
Solar PV	30%	30%	22%	10%	10%	10%	10%	10%	10%	10%

Table A.14: Average annual capacity factors for wind and solar PV technologies for the different temporal representations used in the chronological (C-GEP) and time-slice (TS-GEP) models. Solar PV capacity factors correspond to single-axis tracking PV technology.

	C-GEP	TS-GEP
Wind (existing)	34.6%	36.5%
Wind (new)	38.9%	39.2%
Solar PV (existing and new)	26.3%	27.4%

A.3 Comparison of generation and curtailment between TS-GEP vs C-GEP

We compare the annual generation projections by both GEP models under a hypothetical 50% renewable energy (RE) scenario in Figure A.4.

We also compare the annual curtailment between both GEP models for the 4 RE scenarios in the reference scenario set. Curtailment typically occurs when generation from all sources is higher than the load at a given time, and slow-responding thermal generators cannot be ramped down quickly enough. RE generation is often curtailed in these instances, and therefore we report annual curtailment as a fraction of RE generation. As seen in Figure A.5, the curtailment in 2045 across the 40-60% RE scenarios in the reference scenario set for the TS-GEP model is similar to the curtailment projected by the C-GEP; for the 70% RE scenario, the C-GEP projects much higher curtailment than the TS-GEP (11% vs 6%). The TS-GEP models generation from thermal plants as a continuous variable between zero and the installed nameplate capacity. Therefore, subject to other model constraints, the TS-GEP assumes greater flexibility from the thermal generator fleet than would be available when

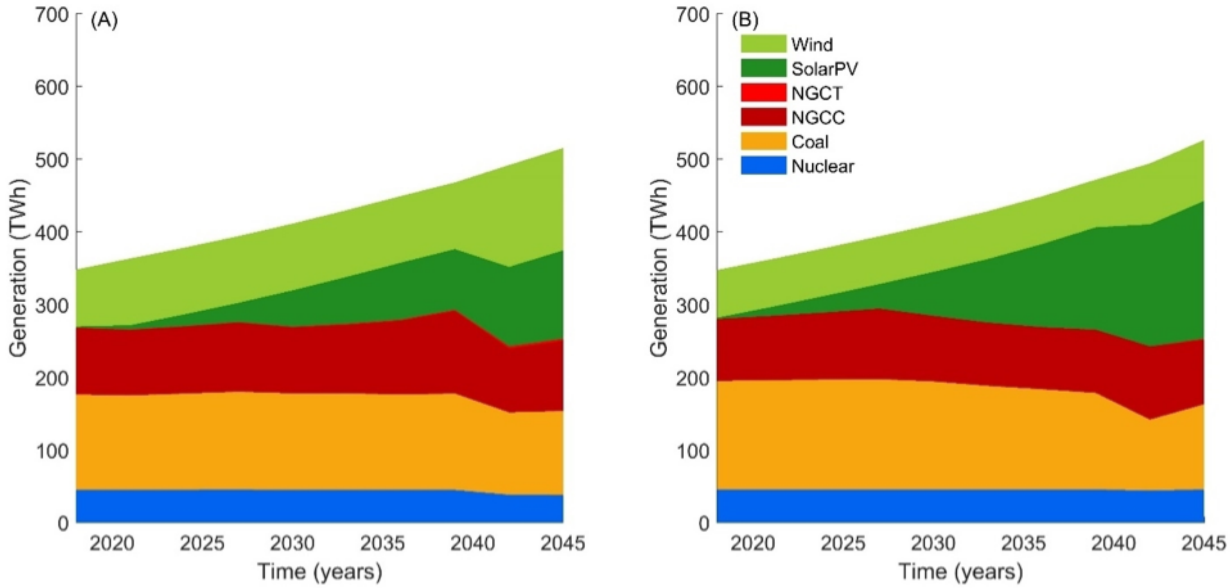

Figure A.4: Comparison of ERCOT generation projections by (A) the chronological model and (B) the time slice model in the 50% RE case.

considering their minimum generation levels (as in the C-GEP). This modeling assumption and the limited representation of temporal variability of load and RE output partly explain why the TS-GEP estimates lower curtailment compared to the C-GEP at higher RE scenarios like 70%.

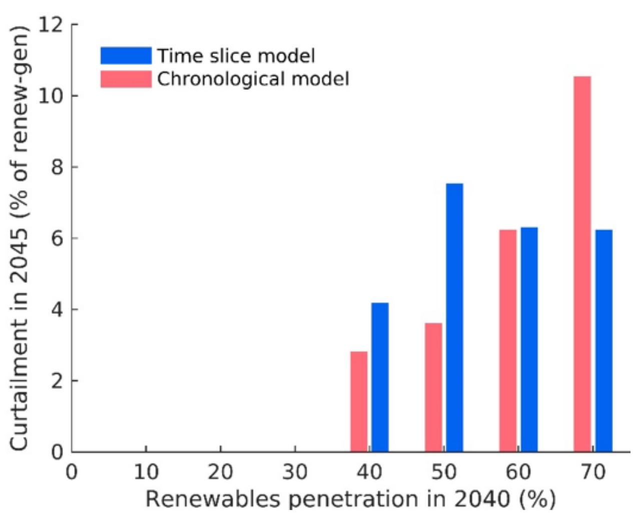

Figure A.5: Curtailment in 2045 estimated by both capacity expansion models across a range of renewables penetration scenarios in the reference set.

A.4 Comparison of outputs between TS-GEP vs. C-GEP using the PCS model

Both GEP models meet RE penetration targets across all the scenarios considered. The PCS model was used to test whether the capacity mix estimated by the GEP models did in fact allow for these targets to be met when considering a full-year simulation of grid operations at hourly resolution for different possible realizations of load and RE outputs. This comparison is presented in the parity plot in Figure A.6. The individual bars refer to the 4 RE scenarios, while the height of the bar represents the range in RE penetration predicted by the PCS for 7 realizations of load and RE capacity factor profiles for each scenario. This figure suggests that that both GEP models are able to meet RE penetration targets reasonably well in these hypothetical model scenarios, given the proximity of the bars to the parity line, although the C-GEP performance is marginally improved compared to the TS-GEP.

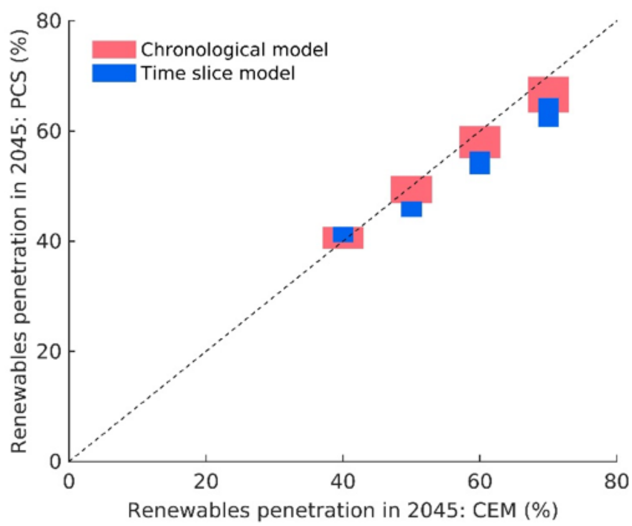

Figure A.6: Comparison of 2045 renewables (RE) penetration in capacity expansion models (TS-GEP and C-GEP) to those in the grid operations model (PCS) for the same capacity mix. The height of each bar corresponds to the range of values obtained from simulating the PCS model for seven different realizations of profiles for load and capacity factors for RE generation (based on 2004-2010 historical data for ERCOT).

In addition, we compare annual thermal generation projected by the GEP models to the

PCS model for consistent capacity mix assumptions across the 4 hypothetical RE scenarios (see Figure A.7). In general, the C-GEP projects annual generation from all thermal sources better than the TS-GEP, as shown in Figure A.7D, specifically, where the C-GEP bars are closer to the parity line. Note that only 3 bars representing nuclear generation from the TS-GEP appear in Figure A.7C because the results of 2 RE scenarios are almost identical and the bars overlap each other.

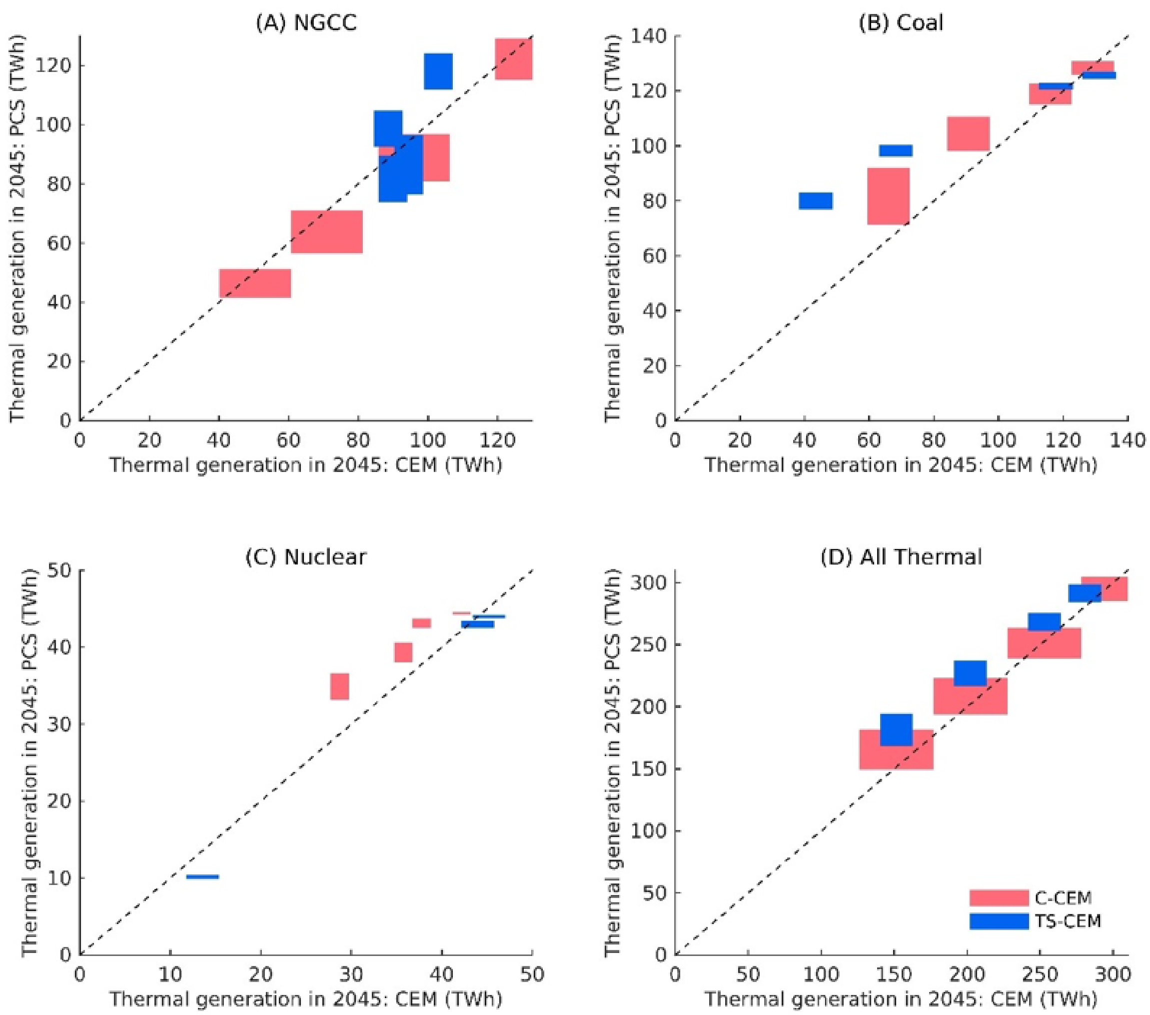

Figure A.7: Comparison of 2045 thermal generation in capacity expansion models (TS-GEP and C-GEP) to those in the grid operations model (PCS) for the same capacity mix. The height of each bar corresponds to the range of values obtained from simulating the PCS model for seven different realizations of profiles for load and capacity factors for renewables generation.

A.5 Additional C-GEP results highlighting impact of number of representative days

Figure A.8 compares the resulting generation projections for NG, solar PV and wind, for each scenario with a different number of sample days used in the C-GEP, under a hypothetical RE 50% target.

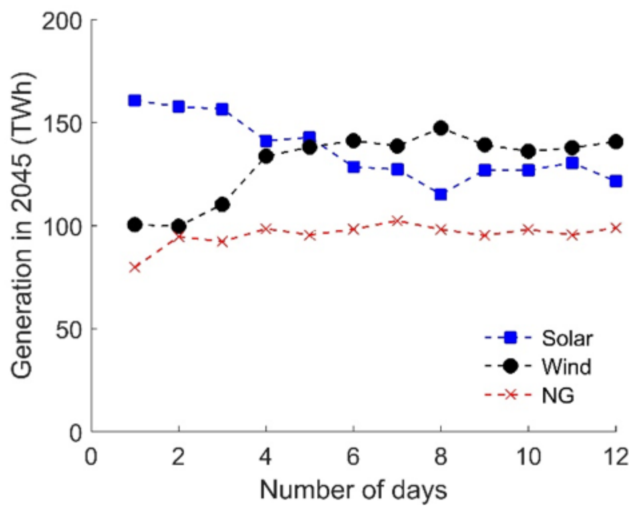

Figure A.8: Generation projections for solar, wind and natural gas in 2045, using the C-GEP under a 50% renewable energy (RE) scenario, varying as a function of the number of sample days selected to represent load and renewables data for annual grid operations. (The L2-norm is used in the k-means clustering approach).

Figure A.9 and Figure A.10 compare the resulting capacity and generation projections for NG, solar PV and wind, for scenarios with a different number of sample days used in the C-GEP, under a RE 50% target, while using the L1-norm as the distance metric in the k-means clustering procedure.

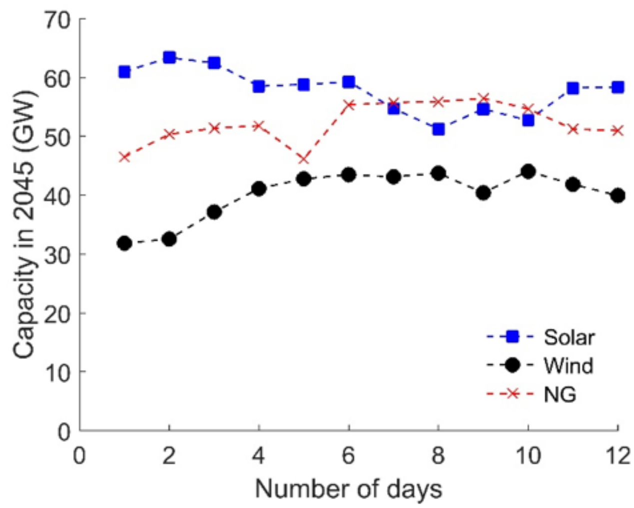

Figure A.9: Capacity projections for solar, wind and natural gas in 2045, using the C-GEP under a 50% renewable energy (RE) scenario, varying as a function of the number of sample days selected to represent load and renewables data for annual grid operations. (The L1-norm is used in the k-means clustering approach).

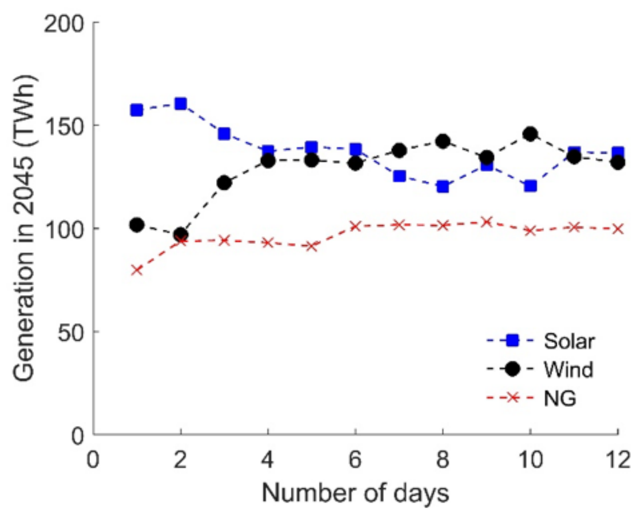

Figure A.10: Generation projections for solar, wind and natural gas in 2045, using the C-GEP under a 50% renewable energy (RE) scenario, varying as a function of the number of sample days selected to represent load and renewables data for one year's operations. (The L1-norm is used in the k-means clustering approach).

Appendix B

Chapter 5 additional material

B.1 Calculated parameters

Regarding the parameters used in equations (5.36)-(5.38), the discount factor in year t, If_t, is calculated from the interest rate, Ir:

$$If_t = \frac{1}{(1+Ir)^t}$$

and the discounted investment cost $DIC_{i,t}$ is given by:

$$DIC_{i,t} = ACC_{i,t} \cdot \left(\sum_{t' \leq \min(LT_i, T^{\text{remain}})} DF_{t'} \right)$$

where the annualized capital cost $ACC_{i,t}$ is given by:

$$ACC_{i,t} = \frac{OCC_{i,t} \cdot Ir}{1 - \frac{1}{(1+Ir)^{LT_i}}}$$

and the remaining time in the horizon T_t^{remain} is defined by $T_t^{\text{remain}} = T - t + 1$.

Appendix C

Chapter 6 additional material

C.1 Detailed MSIP Formulation

The detailed Multistage Stochastic Integer Programming formulation is presented below by equations (C.1)-(C.35). As mentioned before, this is an extension of the MILP model proposed in chapter 5 (Lara et al., 2018a), now including uncertain parameters:

- for the operational uncertainty we have load demand $L_{r,t,d,s,n}$ and renewable capacity factor $Cf_{i,r,t,d,s,n}$ and uncertain parameters drawn from the 2 operational profiles;
- for the strategic uncertainty we have the fuel price $P^{\text{fuel}}_{i,t,n}$ and the carbon tax $Tx^{\text{CO}_2}_{t,n}$, which are considered separately i.e. when fuel price is assumed to be uncertain then carbon tax is assumed to be deterministic, and vice versa.

Note that if an index appears in a summation or next to a $\forall$ symbol without a corresponding set, all elements in that set are assumed.

C.1.1 Energy Balance

Constraint (C.1) ensures that in each sub-period s of representative day d in year t of node n, the sum of instantaneous power $p_{i,r,t,d,s,n}$ generated by generator clusters i in region r plus the difference between the power flow going from regions r' to region r, $p^{\text{flow}}_{r',r,t,d,s,n}$, and the power flowing from region r to regions r', $p^{\text{flow}}_{r,r',t,d,s,n}$, plus the power discharged from all the storage clusters j in region r, $p^{\text{discharge}}_{j,r,t,d,s,n}$, equals the load demand $L_{r,t,d,s,n}$ at that region r, plus the power being charged to the storage clusters j in region r, $p^{\text{charge}}_{j,r,t,d,s,n}$, plus a slack for curtailment of renewable generation $cu_{r,t,d,s,n}$. The distance between regions $D_{r,r'}$ assumes the midpoint for each region, and the transmission loss $T^{\text{loss}}_{r,r'}$ is approximated by a fraction loss per mile.

$$\begin{aligned}&\sum_i (p_{i,r,t,d,s,n}) + \sum_{r'\neq r}\left(p^{\text{flow}}_{r',r,t,d,s,n}\cdot(1-T^{\text{loss}}_{r,r'}\cdot D_{r,r'}) - p^{flow}_{r,r',t,d,s,n}\right) + \sum_j p^{\text{discharge}}_{j,r,t,d,s,n}\\ &= L_{r,t,d,s,n} + \sum_j p^{\text{charge}}_{j,r,t,d,s,n} + cu_{r,t,d,s,n} \qquad \forall\ r, t\in T_n, n, d, s\end{aligned} \tag{C.1}$$

C.1.2 Capacity factor

Constraint (C.2) limits the power outlet $p_{i,r,t,d,s,n}$ of renewable generators to be equal to a fraction $Cf_{i,r,t,d,s,n}$ of the nameplate capacity $Qg^{\text{np}}_{i,r}$ in each sub-period s of representative day d in year t of node n, where $ngo^{\text{rn}}_{i,r,t,n}$ represents the number of renewable generators that are operational in year t of node n. Due to the flexibility in sizes for renewable generators, $ngo^{\text{rn}}_{i,r,t,n}$ is relaxed to be continuous.

$$p_{i,r,t,d,s,n} = Qg^{\text{np}}_{i,r}\cdot Cf_{i,r,t,d,s,n}\cdot ngo^{\text{rn}}_{i,r,t,n} \qquad \forall\ i\in\mathcal{I}^{\text{RN}}_r, r, t\in T_n, n, d, s \tag{C.2}$$

C.1.3 Unit commitment

Constraint (C.3) computes the number of generators that are ON, $u_{i,r,t,d,s,n}$, or in startup, $su_{i,r,t,d,s,n}$, and shutdown, $sd_{i,r,t,d,s,n}$, modes in cluster i in sub-period s of representative day d of year t of node n, and treated as integer variables.

$$u_{i,r,t,d,s,n} = u_{i,r,t,d,s-1,n} + su_{i,r,t,d,s,n} - sd_{i,r,t,d,s,n} \qquad \forall\ i\in\mathcal{I}^{\text{TH}}_r, r, t\in T_n, n, d, s \tag{C.3}$$

C.1.4 Ramping limits

Constraints (C.4)-(C.5) capture the limitation on how fast thermal units can adjust their output power, $p_{i,r,t,d,s,n}$, where $Ru_i^{\max}$ is the maximum ramp-up rate, $Rd_i^{\max}$ is the maximum ramp-down rate, and $Pg_i^{\min}$ is the minimum operating limit Palmintier and Webster (2014).

$$\begin{aligned} p_{i,r,t,d,s,n} - p_{i,r,t,d,s-1,n} \leq \quad & Ru_i^{\max} \cdot Hs \cdot Qg_{i,r}^{\text{np}} \cdot (u_{i,r,t,d,s,n} - su_{i,r,t,d,s,n}) \\ & + \max\left(Pg_i^{\min}, Ru_i^{\max} \cdot Hs\right) \cdot Qg_{i,r}^{\text{np}} \cdot su_{i,r,t,d,s,n} \\ & \forall\ i \in \mathcal{I}_r^{\text{TH}}, r, t \in T_n, n, d, s \end{aligned} \tag{C.4}$$

$$\begin{aligned} p_{i,r,t,d,s-1,n} - p_{i,r,t,d,s,n} \leq \quad & Rd_i^{\max} \cdot Hs \cdot Qg_{i,r}^{\text{np}} \cdot (u_{i,r,t,d,s,n} - su_{i,r,t,d,s,n}) \\ & + \max\left(Pg_i^{\min}, Rd_i^{\max} \cdot Hs\right) \cdot Qg_{i,r}^{\text{np}} \cdot sd_{i,r,t,d,s,n} \\ & \forall\ i \in \mathcal{I}_r^{\text{TH}}, r, t \in T_n, n, d, s \end{aligned} \tag{C.5}$$

C.1.5 Operating limits

Constraints (C.6)-(C.7) specify that each thermal generator is either OFF and outputting zero power, or ON and running within the operating limits $Pg_i^{\min} \cdot Qg_{i,r}^{\text{np}}$ and $Qg_{i,r}^{\text{np}}$. The variable $u_{i,r,t,d,s,n}$ (integer variable) represents the number of generators that are ON in cluster $i \in \mathcal{I}_r^{\text{TH}}$ at the time period t of node n, representative day d, and sub-period s. Note that constraint (C.7) is modified in order to capture the need for generators to run below the maximum considering operating reserves, where $q_{i,r,t,d,s}^{\text{spin}}$ is a variable representing the spinning reserve capacity.

$$u_{i,r,t,d,s,n} \cdot Pg_i^{\min} \cdot Qg_{i,r}^{\text{np}} \leq p_{i,r,t,d,s,n} \qquad \forall\ i \in \mathcal{I}_r^{\text{TH}}, r, t \in T_n, n, d, s \tag{C.6}$$

$$p_{i,r,t,d,s,n} + q_{i,r,t,d,s,n}^{\text{spin}} \leq u_{i,r,t,d,s,n} \cdot Qg_{i,r}^{\text{np}} \qquad \forall\ i \in \mathcal{I}_r^{\text{TH}}, r, t \in T_n, n, d, s \tag{C.7}$$

C.1.6 Total operating reserve

Constraint (C.8) dictates that the total spinning reserve, $q_{i,r,t,d,s,n}^{\text{spin}}$, plus quick-start reserve, $q_{i,r,t,d,s,n}^{\text{Qstart}}$, must exceed the minimum operating reserve, $Op^{\min}$, which is a percentage of the load $L_{r,t,d,s,n}$ in a reserve sharing region r at each sub-period s.

$$\sum_{i \in \mathcal{I}_r^{\mathrm{TH}}} \left(q_{i,r,t,d,s,n}^{\mathrm{spin}} + q_{i,r,t,d,s,n}^{\mathrm{Qstart}} \right) \geq Op^{\min} \cdot L_{r,t,d,s,n} \qquad \forall\ r, t \in T_n, n, d, s \tag{C.8}$$

C.1.7 Total spinning reserve

Constraint (C.9) specifies that the total spinning reserve $q_{i,r,t,d,s,n}^{\mathrm{spin}}$ must exceed the minimum spinning reserve, $Spin^{\min}$, which is a percentage of the load $L_{r,t,d,s,n}$ in a reserve sharing region r at each sub-period s.

$$\sum_{i \in \mathcal{I}_r^{\mathrm{TH}}} q_{i,r,t,d,s,n}^{\mathrm{spin}} \geq Spin^{\min} \cdot L_{r,t,d,s,n} \qquad \forall\ r, t \in T_n, n, d, s \tag{C.9}$$

C.1.8 Maximum spinning reserve

Constraint (C.10) states that the maximum fraction of capacity of each generator cluster that can contribute to spinning reserves is given by $Frac_i^{\mathrm{spin}}$, which is a fraction of the nameplate capacity $Qg_{i,r}^{\mathrm{np}}$.

$$q_{i,r,t,d,s,n}^{\mathrm{spin}} \leq u_{i,r,t,d,s,n} \cdot Qg_{i,r}^{\mathrm{np}} \cdot Frac_i^{\mathrm{spin}} \qquad \forall\ i \in \mathcal{I}_r^{\mathrm{TH}}, r, t \in T_n, n, d, s \tag{C.10}$$

C.1.9 Maximum quick-start reserve

Constraint (5.11) dictates that the maximum fraction of the capacity of each generator cluster that can contribute to quick-start reserves is given by $Frac_i^{\mathrm{Qstart}}$ (fraction of the nameplate capacity $Qg_{i,r}^{\mathrm{np}}$), and that quick-start reserves can only be provided by the generators that are OFF, i.e., not active. Here the integer variable $ngo_{i,r,t,n}^{\mathrm{th}}$ represents the number of thermal generators that are operational (i.e., installed and ready to operate) at year t of node n.

$$q_{i,r,t,d,s,n}^{\mathrm{Qstart}} \leq (ngo_{i,r,t,n}^{\mathrm{th}} - u_{i,r,t,d,s,n}) \cdot Qg_{i,r}^{\mathrm{np}} \cdot Frac_i^{\mathrm{Qstart}} \qquad \forall\ i \in \mathcal{I}_r^{\mathrm{TH}}, r, t \in T_n, n, d, s \tag{C.11}$$

C.1.10 Planning reserve requirement

Constraint (C.12) ensures that the operating capacity is greater than or equal to the annual peak load $L_t^{\max}$, plus a predefined fraction of reserve margin $R_t^{\min}$ of the annual peak load $L_t^{\max}$. Due to the due to the renewables inability to control dispatch and the uncertainty of the output, only a fraction of their nameplate capacity, referred to as the capacity value Q_i^{v} counts towards the planning reserve requirement.

$$\begin{aligned} &\sum_{i\in\mathcal{I}_r^{\mathrm{RN}}}\sum_{r}\left(Qg_{i,r}^{\mathrm{np}}\cdot Q_i^{\mathrm{v}}\cdot ngo_{i,r,t,n}^{\mathrm{rn}}\right)+\sum_{i\in\mathcal{I}_r^{\mathrm{TH}}}\sum_{r}\left(Qg_{i,r}^{\mathrm{np}}\cdot ngo_{i,r,t,n}^{\mathrm{th}}\right) \\ &\geq (1+R_t^{\min})\cdot L_t^{\max} \qquad \forall\, t\in T_n, n \end{aligned} \tag{C.12}$$

C.1.11 Minimum annual renewable generation requirement

Constraint (C.13) ensures that, in case of policy mandates, the renewable generation quota target, $RN_t^{\min}$, which is a fraction of the energy demand $ED_{t,n}$, is satisfied. If not, i.e, if there is a deficit $def_{t,n}^{\mathrm{rn}}$ from the quota, this is subjected to a penalty that is included later in the objective function.

$$\begin{aligned} &\sum_{d}\sum_{s}\left[W_d\cdot Hs\cdot\left(\sum_{i\in\mathcal{I}_r^{\mathrm{RN}}}\sum_{r}p_{i,r,t,d,s,n}-cu_{r,t,d,s,n}\right)\right]+def_{,n}^{\mathrm{rn}}t \\ &\geq RN_t^{\min}\cdot ED_t \qquad \forall\, t\in T_n, n \end{aligned} \tag{C.13}$$

Here W_d represents the weight of the representative day d, Hs is the length of the sub-period, $cu_{r,t,d,s,n}$ is the curtailment of renewable generation, and $ED_{t,n}$ represent the energy demand in year t of node n:

$$ED_{t,n}=\sum_{r}\sum_{d}\sum_{s}\left(W_d\cdot Hs\cdot L_{r,t,d,s,n}\right)$$

C.1.12 Maximum yearly installation

Constraints (C.14)-(C.15) limit the yearly installation per generation type in each region r to an upper bound $Q_{i,t}^{\text{inst,UB}}$ in MW/year. Here $ngb_{i,r,t,n}^{\text{rn}}$ and $ngb_{i,r,t,n}^{\text{th}}$ represent the number of renewable and thermal generators built in region r in year t of node n, respectively. Note that due to the flexibility in sizes for renewable generators, $ngb_{i,r,t,n}^{\text{rn}}$ is relaxed to be continuous.

$$\sum_r ngb_{i,r,t,n}^{\text{rn}} \leq Q_{i,t}^{\text{inst,UB}}/Qg_{i,r}^{\text{np}} \qquad \forall\ i \in \mathcal{I}_r^{\text{Rnew}}, t \in T_n, n \tag{C.14}$$

$$\sum_r ngb_{i,r,t,n}^{\text{th}} \leq Q_{i,t}^{\text{inst,UB}}/Qg_{i,r}^{\text{np}} \qquad \forall\ i \in \mathcal{I}_r^{\text{Tnew}}, t \in T_n, n \tag{C.15}$$

C.1.13 Balance of generators

Concerning renewable generator clusters, we define a set of constraints (C.16)-(C.17) to compute the number of generators in cluster i that are ready to operate $ngo_{i,r,t,n}^{\text{rn}}$, taking into account the generators that were already existing at the beginning of the planning horizon $Ng_{i,r}^{\text{Rold}}$, the generators built $ngb_{i,r,t,n}^{\text{rn}}$, and the generators retired $ngr_{i,r,t,n}^{\text{rn}}$ at year t of node n. It is important to highlight that we assume *no lead time* between the decision to build/install a generator and the moment it can begin producing electricity.

$$ngo_{i,r,t,n}^{\text{rn}} = Ng_{i,r}^{\text{Rold}} + ngb_{i,r,t,n}^{\text{rn}} - ngr_{i,r,t,n}^{\text{rn}} \qquad \forall\ i \in \mathcal{I}_r^{\text{RN}}, r, t = 1, n = 1 \tag{C.16}$$

$$ngo_{i,r,t,n}^{\text{rn}} = ngo_{i,r,t-1,P(n)}^{\text{rn}} + ngb_{i,r,t,n}^{\text{rn}} - ngr_{i,r,t,n}^{\text{rn}} \qquad \forall\ i \in \mathcal{I}_r^{\text{RN}}, r, t \in T_n, n > 1 \tag{C.17}$$

As aforementioned, due to the flexibility in sizes for renewable generators, $ngo_{i,r,t,n}^{\text{rn}}$, $ngb_{i,r,t,n}^{\text{rn}}$, and $ngr_{i,r,t,n}^{\text{rn}}$ are relaxed to be continuous. Note that $ngb_{i,r,t,n}^{\text{rn}}$ for $i \in \mathcal{I}_r^{\text{Rold}}$ is fixed to zero in all time periods, i.e., the clusters of existing renewable generators cannot have any new additions during the time horizon considered.

We also define constraint (C.18) to enforce the renewable generators that reached the end of their lifetime to either retire, $ngr_{i,r,t,n}^{\text{rn}}$, or have their life extended, $nge_{i,r,t,n}^{\text{rn}}$. $Ng_{i,r,t}^{\text{r}}$

is a parameter that represents the number of old generators (i.e., $i \in \mathcal{I}_r^{\text{old}}$) that reached the end of their lifetime, LT_i, at year t.

$$Ng^{\text{r}}_{i,r,t} = ngr^{\text{rn}}_{i,r,t,n} + nge^{\text{rn}}_{i,r,t,n} \qquad \forall\, i \in \mathcal{I}_r^{\text{Rold}}, r, t \in T_n, n \tag{C.18}$$

Concerning thermal generator clusters, we define a set of constraints (C.19)-(C.20) to compute the number of generators in cluster i that are ready to operate $ngo^{\text{th}}_{i,r,t,n}$, taking into account the generators that were already existing at the beginning of the planning horizon $Ng^{\text{Told}}_{i,r}$, the generators built $ngb^{\text{th}}_{i,r,t,n}$, and the generators retired $ngr^{\text{th}}_{i,r,t,n}$ at year t of node n.

$$ngo^{\text{th}}_{i,r,t,n} = Ng^{\text{Told}}_{i,r} + ngb^{\text{th}}_{i,r,t,n} - ngr^{\text{th}}_{i,r,t,n} \qquad \forall\, i \in \mathcal{I}_r^{\text{TH}}, r, t = 1, n = 1 \tag{C.19}$$

$$ngo^{\text{th}}_{i,r,t,n} = ngo^{\text{th}}_{i,r,t-1,P(n)} + ngb^{\text{th}}_{i,r,t,n} - ngr^{\text{th}}_{i,r,t,n} \qquad \forall\, i \in \mathcal{I}_r^{\text{TH}}, r, t \in T_n, n > 1 \tag{C.20}$$

Note that $ngb^{\text{th}}_{i,r,t,n}$ for $i \in \mathcal{I}_r^{\text{Told}}$ is fixed to zero in all time periods, i.e., the clusters of existing thermal generators cannot have any new additions during the time horizon considered.

We also define constraint (C.21) to enforce the thermal generators that reached the end of their lifetime to either retire, $ngr^{\text{th}}_{i,r,t,n}$, or have their life extended $nge^{\text{th}}_{i,r,t,n}$.

$$Ng^{\text{r}}_{i,r,t} = ngr^{\text{th}}_{i,r,t,n} + nge^{\text{th}}_{i,r,t,n} \qquad \forall\, i \in \mathcal{I}_r^{\text{Told}}, r, t \in T_n, n \tag{C.21}$$

Finally, we have constraint (C.22) that ensures that only installed generators can be in operation:

$$u_{i,r,t,d,s,n} \le ngo^{\text{th}}_{i,r,t,n} \qquad \forall\, i \in \mathcal{I}_r^{\text{Tnew}}, r, t \in T_n, n, d, s \tag{C.22}$$

C.1.14 Storage

The energy storage devices are assumed to be ideal and generic Pozo et al. (2014). Constraints (C.23)-(C.24) compute the number of storage units that are ready to operate $nso_{j,r,t,n}$, taking into account the storage units already existing at the beginning of the planning horizon $Ns_{j,r}$ and the ones built $nsb_{j,r,t,n}$ at year t of node n. Due to the flexibility in sizes for storage units, $nso_{j,r,t,n}$ and $nsb_{j,r,t,n}$ are relaxed to be continuous.

$$nso_{j,r,t,n} = Ns_{j,r} + nsb_{j,r,t,n} \qquad \forall\, j, r, t = 1, n = 1 \tag{C.23}$$

$$nso_{j,r,t,n} = nso_{j,r,t-1,P(n)} + nsb_{j,r,t,n} \qquad \forall\, j, r, t \in T_n, n > 1 \tag{C.24}$$

Constraints (C.25) and (C.26) establish that the power charge, $p^{\text{charge}}_{j,r,t,d,s,n}$, and discharge, $p^{\text{discharge}}_{j,r,t,d,s,n}$, of the storage units in cluster j, $nso_{j,r,t,n}$, has to be within the operating limits: $Charge^{\min}_j$ and $Charge^{\max}_j$, and $Discharge^{\min}_j$ and $Discharge^{\min}_j$, respectively.

$$Charge^{\min}_j \cdot nso_{j,r,t,n} \leq p^{\text{charge}}_{j,r,t,d,s,n} \leq Charge^{\max}_j \cdot nso_{j,r,t,n} \qquad \forall\, j, r, t \in T_n, n, d, s \tag{C.25}$$

$$Discharge^{\min}_j \cdot nso_{j,r,t,n} \leq p^{\text{discharge}}_{j,r,t,d,s,n} \leq Discharge^{\max}_j \cdot nso_{j,r,t,n} \qquad \forall\, j, r, t \in T_n, n, d, s \tag{C.26}$$

Constraint (C.27) specifies that the energy storage level, $p^{\text{level}}_{j,r,t,d,s,n}$, for the storage units in cluster j, $nso_{j,r,t,n}$ has to be within the storage capacity limits $Storage^{\min}_j$ and $Storage^{\max}_j$.

$$Storage^{\min}_j \cdot nso_{j,r,t,n} \leq p^{\text{level}}_{j,r,t,d,s} \leq Storage^{\max}_j \cdot nso_{j,r,t,n} \qquad \forall\, j, r, t \in T_n, n, d, s \tag{C.27}$$

Constraints (C.28) and (C.29) show the power balance in the storage units. The state of charge $p^{\text{level}}_{j,r,t,d,s,n}$ at the end of sub-period s depends on the previous state of charge $p^{\text{level}}_{j,r,t,d,s-1,n}$, and the power charged $p^{\text{charge}}_{j,r,t,d,s,n}$ and discharged $p^{\text{discharge}}_{j,r,t,d,s,n}$ at sub-period s. The symbols η^{charge}_j and $\eta^{\text{discharge}}_j$ represent the charging and discharging efficiencies, respectively. For the first hour of the day d in year t of node n, the previous state of charge (i.e., $s = 0$) is the variable $p^{\text{level},0}_{j,r,t,d,n}$.

$$p^{\text{level}}_{j,r,t,d,s,n} = p^{\text{level}}_{j,r,t,d,s-1,n} + \eta^{\text{charge}}_j \cdot p^{\text{charge}}_{j,r,t,d,s} + p^{\text{discharge}}_{j,r,t,d,s,n}/\eta^{\text{discharge}}_j \qquad \forall\, j, r, t \in T_n, n, d, s > 1 \tag{C.28}$$

$$p^{\text{level}}_{j,r,t,d,s,n} = p^{\text{level},0}_{j,r,t,d,n} + \eta^{\text{charge}}_j \cdot p^{\text{charge}}_{j,r,t,d,s,n} + p^{\text{discharge}}_{j,r,t,d,s,n}/\eta^{\text{discharge}}_j \qquad \forall\, j, r, t \in T_n, n, d, s = 1 \tag{C.29}$$

Constraints (C.30) and (C.31) force the storage units to begin $p^{\text{level},0}_{j,r,t,d,s}$ and end $p^{\text{level}}_{j,r,t,d,s=S,n}$ each day d of year t with 50% of their maximum storage $Storage^{\max}_j$. This is a heuristic to attach carryover storage level form one representative day to the next Liu et al. (2018).

$$p^{\text{level},0}_{j,r,t,d,n} = 0.5 \cdot Storage^{\max}_j \cdot nso_{j,r,t,n} \qquad \forall\ j, r, t \in T_n, n, d \tag{C.30}$$

$$p^{\text{level}}_{j,r,t,d,s,n} = 0.5 \cdot Storage^{\max}_j \cdot nso_{j,r,t,n} \qquad \forall\ j, r, t \in T_n, n, d, s = S \tag{C.31}$$

C.1.15 Objective function

The objective of this model is to minimize the *expected* net present cost, Φ, over the planning horizon, which includes operating costs Φ^{opex}, investment costs Φ^{capex}, and potential penalties Φ^{PEN} for not meeting the the targets on renewables.

$$\min \quad \Phi = \sum_{n \in \mathcal{T}} prob_n \cdot \sum_{t \in T_n} \left(\Phi^{\text{opex}}_{t,n} + \Phi^{\text{capex}}_{t,n} + \Phi^{\text{PEN}}_{t,n} \right) \tag{C.32}$$

The operating expenditure, $\Phi^{\text{opex}}_{t,n}$, comprises the variable $VOC_{i,t}$ and fixed $FOC_{i,t}$ operating costs, as well as fuel cost $P^{\text{fuel}}_{i,t,n}$ per heat rate HR_i, carbon tax $Tx^{\text{CO}_2}_{t,n}$ for CO_2 emissions $EF^{CO_2}_i$, and start-up cost (variable cost $P^{\text{fuel}}_{i,t,n}$ that depends on the amount of fuel burned for startup F^{start}_i, and fixed cost C^{start}_i). Both $P^{\text{fuel}}_{i,t,n}$ and $Tx^{\text{CO}_2}_{t,n}$ are potential strategic uncertain parameters, hence are indexed by node n.

$$\begin{aligned}
\Phi^{\text{opex}}_{t,n} = \quad & If_t \cdot \Bigg[\sum_d \sum_s W_d \cdot hs \cdot \left(\sum_i \sum_r (VOC_{i,t} + P^{\text{fuel}}_{i,t,n} \cdot HR_i + Tx^{\text{CO}_2}_{t,n} \cdot EF^{CO_2}_i \cdot HR_i) \cdot p_{i,r,t,d,s,n} \right) \\
& + \left(\sum_{i \in \mathcal{I}^{\text{RN}}_r} \sum_r FOC_{i,t} \cdot Qg^{\text{np}}_{i,r} \cdot ngo^{\text{rn}}_{i,r,t,n} \right) + \left(\sum_{i \in \mathcal{I}^{\text{TH}}_r} \sum_r FOC_{i,t} \cdot Qg^{\text{np}}_{i,r} \cdot ngo^{\text{th}}_{i,r,t,n} \right) \\
& + \sum_{i \in \mathcal{I}^{\text{TH}}_r} \sum_r \sum_d \sum_s W_d \cdot Hs \cdot su_{i,r,t,d,s,n} \cdot Qg^{\text{np}}_{i,r} \\
& \cdot \left(F^{\text{start}}_i \cdot P^{\text{fuel}}_{i,t,n} + F^{\text{start}}_i \cdot EF^{CO_2} \cdot Tx^{\text{CO}_2}_{t,n} + C^{\text{start}}_i \right) \Bigg] \qquad \forall\ n \in \mathcal{T},\ t \in T_n
\end{aligned} \tag{C.33}$$

The capital expenditure, $\Phi^{\text{capex}}_{t,n}$, includes the amortized cost of acquiring new generators, $DIC_{i,t}$, new storage devices, $SIC_{j,t}$, and the amortized cost of extending the life of generators that reached their expected lifetime. The latter is assumed to be a fraction LE_i of the investment cost, $DIC_{i,t}$, in a new generator with the same or equivalent generation technology. In this framework, the investment cost takes into account the remaining value at the end of the time horizon by considering the annualized capital cost and multiplying it by the number of years remaining in the planning horizon at the time of installation to calculate the $DIC_{i,t}$.

$$\begin{aligned}
\Phi_{t,n}^{\text{capex}} = \quad & If_t \cdot \Bigg[\sum_{i \in \mathcal{I}_r^{\text{Rnew}}} \sum_r DIC_{i,t} \cdot CC_i^{\text{m}} \cdot Qg_{i,r}^{\text{np}} \cdot ngb_{i,r,t,n}^{\text{rn}} \\
& + \sum_{i \in \mathcal{I}_r^{\text{Tnew}}} \sum_r DIC_{i,t} \cdot CC_i^{\text{m}} \cdot Qg_{i,r}^{\text{np}} \cdot ngb_{i,r,t,n}^{\text{th}} \\
& + \sum_j \sum_r SIC_{j,t} \cdot Storage_j^{\max} \cdot nsb_{j,r,t,n} \\
& + \sum_{i \in \mathcal{I}_r^{\text{RN}}} \sum_r DIC_{i,t} \cdot LE_i \cdot Qg_{i,r}^{\text{np}} \cdot nge_{i,r,t,n}^{\text{rn}} \\
& + \sum_{i \in \mathcal{I}_r^{\text{TH}}} \sum_r DIC_{i,t} \cdot LE_i \cdot Qg_{i,r}^{\text{np}} \cdot nge_{i,r,t,n}^{\text{th}} \Bigg] \qquad \forall\, n \in \mathcal{T},\ t \in T_n
\end{aligned} \tag{C.34}$$

The capital multiplier CC_i^{m} associated with new generator clusters is meant to account for differences in depreciation schedules applicable to each technology, with higher values being indicative of slower depreciating schedule and vice versa.

Lastly, the penalty cost, $\Phi_{t,n}^{\text{PEN}}$, includes the potential fines for not meeting the renewable energy quota, PEN_t^{rn}, and curtailing the renewable generation.

$$\Phi_{t,n}^{\text{PEN}} = \quad If_t \cdot \left(PEN_t^{\text{rn}} \cdot def_{t,n}^{\text{rn}} + PEN^c \cdot \sum_r \sum_d \sum_s cu_{r,t,d,s,n} \right) \tag{C.35}$$