UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Assessment and Methods for Supply-Following Loads in Modern Electricity Grids with Deep Renewables Penetration

Permalink
https://escholarship.org/uc/item/5x59v46p

Author
Taneja, Jayant Kumar

Publication Date
2013

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Assessment and Methods for Supply-Following Loads in Modern Electricity Grids with Deep Renewables Penetration

by

Jayant Kumar Taneja

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David E. Culler, Chair
Professor Randy H. Katz
Assistant Professor Duncan S. Callaway

Fall 2013

Assessment and Methods for Supply-Following Loads in Modern Electricity Grids with Deep Renewables Penetration

Copyright 2013
by
Jayant Kumar Taneja

Abstract

Assessment and Methods for Supply-Following Loads in Modern Electricity Grids with Deep Renewables Penetration

by

Jayant Kumar Taneja

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor David E. Culler, Chair

Electricity is an indispensable commodity to modern society, yet it is delivered via a grid architecture that remains largely unchanged over the past century. A host of factors are conspiring to topple this dated yet venerated design: developments in renewable electricity generation technology, policies to reduce greenhouse gas emissions, and advances in information technology for managing energy systems. Modern electric grids are emerging as complex distributed systems in which a portfolio of power generation resources, often incorporating fluctuating renewable resources such as wind and solar, must be managed dynamically to meet uncontrolled, time-varying demand. Uncertainty in both supply and demand makes control of modern electric grids fundamentally more challenging, and growing portfolios of renewables exacerbate the challenge.

We study three electricity grids: the state of California, the province of Ontario, and the country of Germany. To understand the effects of increasing renewables, we develop a methodology to scale renewables penetration. Analyzing these grids yields key insights about rigid limits to renewables penetration and their implications in meeting long-term emissions targets. We argue that to achieve deep penetration of renewables, the operational model of the grid must be inverted, changing the paradigm from load-following supplies to supply-following loads.

To alleviate the challenge of supply-demand matching on deeply renewable grids, we first examine well-known techniques, including altering management of existing supply resources, employing utility-scale energy storage, targeting energy efficiency improvements, and exercising basic demand-side management. Then, we create several instantiations of supply-following loads – including refrigerators, heating and cooling systems, and laptop computers – by employing a combination of sensor networks, advanced control techniques, and enhanced energy storage. We examine the capacity of each load for supply-following and study the behaviors of populations of these loads, assessing their potential at various levels of deployment throughout the California electricity grid. Using combinations of supply-following strategies, we can reduce peak natural gas generation by 19% on a model of the

California grid with 60% renewables. We then assess remaining variability on this deeply renewable grid incorporating supply-following loads, characterizing additional capabilities needed to ensure supply-demand matching in future sustainable electricity grids.

Contents

List of Figures

List of Tables

Acknowledgments

Though it was a grueling, humbling, and ultimately gratifying individual challenge to get to this point, I "stood on the shoulders of giants," and must mention and thank the many people and organizations that made this work possible by helping me along the way.

First, I would like to thank my parents, Usha and Dinesh Taneja, who have made enormous sacrifices to provide me with the opportunities I have had in life. Simply put, there is no way to repay this debt. I also would like to thank my brother Raj, sister-in-law Patty, and nephews Aidan and Sam for providing a light at the end of the tunnel each semester – they have been a bundle of vitality, always ready to receive me whenever I could visit. Raj was especially sagely in discussing the trials and tribulations of graduate school, helping me through some very tough times along the way.

Next, I would like to thank my extended family, mostly in India, who continue to embody the concept of unconditional love even after I disappeared from their lives for so many years. I am incredibly lucky to have such pure and wonderful people as part of my life.

So much of my academic and personal growth is due to my graduate advisor, David Culler. In knowing when to press and when to pitch in, which directions to chase and which to avoid, and how high to set the bar, David has been a strong force in my development as a researcher. I truly believe that I have learned the craft of academic research in systems from one of its masters, and cannot thank him enough for his confidence in me that allowed me to build that confidence in myself.

There are many other professors and researchers at Berkeley and elsewhere that have provided assistance, collaboration, instruction, and advice along the way. I would like to thank Randy Katz and Kara Nelson for their support and collaboration through the years. Catherine Rosenberg was a surprise collaborator that has been extremely valuable for enhancing the rigor of my work. Working with Todd Dawson was critical for developing my master's thesis. Duncan Callaway, Dave Auslander, and Costas Spanos provided excellent feedback in guiding my dissertation work. Steven Lanzisera, Rich Brown, and Ken Lutz were fantastic research collaborators. Steve Bibyk and Jose Gutierrez were critical early influences that led me to pursue a graduate career in my field, for which I am very grateful.

My graduate school peers are likely the most talented group of people with whom I will ever work. For the many long hours of discussion, the moments of release during and after difficult work, and the extensive collaboration (many are co-authors), I thank them deeply, and mention them by name: Prabal Dutta (the best mentor one could ask for), Stephen Dawson-Haggerty, Anil Aswani, Jorge Ortiz, Andrew Krioukov, Fred Jiang, Jaein Jeong, Sukun Kim, Prashanth Mohan, and Arka Bhattacharya.

A special thank you goes to the support staff who made this research possible. Albert Goto is one of the nicest, most helpful people I have ever met, and I have sincere gratitude for all of the late nights, long discussions, and crucial assistance he provided throughout my long tenure at Berkeley. Domenico Caramagno and Scott McNally provided unprecedented access to the building systems they are responsible for, enabling the research of our group to exist and thrive.

I can attribute a great deal of my sanity to the extremely supportive and deep friendships built while I lived in Berkeley and San Francisco. The Beer of the Week club, the Mountain Man Competition, Team Us dinners, and countless other get-togethers provide the context for this amazing period of my life, and the people who made life wonderful hold a dear place in my heart. A special thank you goes to my roommates at Brokeback Milvia, the Copa Colusa, and 995 – I slept on the couch a lot.

I also have a deep appreciation for the public education system, of which I am a product. The resources and full support of the University of California, Berkeley, and The Ohio State University have empowered me to succeed. There is no institution like Berkeley in the world, and I am grateful that I was able to experience graduate school there. I also want to thank the many teachers along the way, from elementary school onwards, whose hard work shaped me into what I am today. Additionally, I would like to acknowledge the public funding sources that enabled the graduate research of myself and so many others, including the NSF and the DoD, especially via the NDSEG Fellowship.

Last, I would like to thank my wife-to-be, Emily, who has been a sounding board, motivator, escape companion, stabilizing force, and inspiration during the latter half of my graduate school career. I cannot say where I would be without her, so I will not try. I can only say that I love her no matter what and that I am incredibly happy and lucky to be taking this journey together.

Chapter 1

Introduction

The electric grid has been called the most important invention of the Industrial Age. In the analogy of a supply chain, it is akin to a massive distribution system, where hundreds of suppliers are connected to millions of customers – each with individual, time-varying needs – in essentially real-time and almost exclusively without the benefit of inventory or warehouses. Classical grid design consists of centralized, large-scale, and primarily fossil-fueled generation sources; long-distance, alternating current (AC) transmission networks; and immense numbers of widely distributed and heterogenous electric loads operating entirely unaware of the activities on the rest of the grid. This design has held true for over a century since the emergence of electricity as a widespread human energy source in the latter half of the 19th century. It is said that if Alexander Graham Bell were alive today, he would surely be astonished with the innovations to his groundbreaking invention – mobile telephones, streaming video, and essentially ubiquitous connectivity affect humanity every day. On the other hand, if Thomas Edison were alive, he would look at the electric power grid and easily recognize the building blocks that he helped design. This adage serves two purposes: first, to showcase the rapid pace of development in information technology, and second, to highlight that information technology has yet to significantly penetrate the design of electricity grids.

Electric grids in the modern era have now reached an inflection point. In many cases, localities have adopted renewables portfolio standards, policies to produce certain percentages of electric energy from renewable sources by some future year (*e.g.*, California aims for renewables to comprise 33% of its energy mix by 2020 [112]). In fact, 36 of the 50 U.S. states have set goals for an RPS of anywhere from 10 to 25% of total energy consumed [37]. The proliferation of these and other policies to reduce emissions of greenhouse gases in electricity generation, coupled with volatility in fuel markets and the emergence of information technology for managing massively distributed systems, augurs an epoch of massive transformation for the electricity industry. This wave of upheaval has resulted in a self-reinforcing groundswell of ever-cheaper renewable electricity sources, chiefly terrestrial wind farms as well as utility-scale and distributed solar power systems, leading to electricity generation mixes that are a vast departure from grids today.

Following the course of this trend, grids with deep renewables penetration present a

family of new challenges and opportunities in maintaining a reliable electricity system. At core, the challenges arise from the shift from primarily modulated, dispatchable supply, where energy generation can be scheduled and controlled, to primarily uncontrolled, non-dispatchable supply, and the resultant opportunities from the increased intelligence and communication needed to allow the energy network to function *as a system*. In this type of system, maintaining the balance between supply and demand transitions from a completely centralized challenge managed by electricity providers to a distributed one involving both providers and consumers. This evolution to a more integrated, cooperative network of sources and loads requires a far more extensive data overlay of control and management to maintain the dynamic match between electricity supply and demand.

There are many ongoing efforts in technologies for improving the balance of supply and demand, primarily in energy storage and demand response. Existing grid-scale energy storage, consisting almost entirely of pumped hydroelectric storage, suffers from unfavorable economics, resulting in a lack of scale – there are only 22 GW of energy storage in the U.S., an amount roughly equivalent to 2.5% of national generation capacity. Though there is much promising research in this area, there is a clear lack of demonstrated technologies that meet the cost and scale requirements of grid-scale operation [89]. Demand response aims to address the balance by actuating the demand-side; a utility secures contracts with larger consumers of electricity (*e.g.*, an aluminum smelting plant) to request curtailment of electricity consumption on-demand. Though there are efforts to enroll more customers and automate the DR process [12, 113], most demand response today involves a small number of large electricity consumers with sub-hourly metering infrastructure [28].

Despite the availability of these technologies to help manage the match between supply and demand, there are high-profile examples of grids with significant renewables penetration struggling to balance supply and demand and maintain electrical reliability. For example, the Texas grid had seven declared power emergencies in the summer of 2011, resulting from a combination of increased electricity demand, poor output from the large contingent of wind turbines in Texas (the most in the U.S.), and routine maintenance to other generation facilities [91]. In two of those cases, a Level 2 emergency was declared, where the demand response capacity of large industrial customers was called in to avoid rolling blackouts. As the state continues to deploy further wind generation and reserve margins of extra power available continue to shrink, the state is likely to face further electricity emergencies and potential blackouts. This anecdote serves to highlight the concept of *limits to renewables penetration*, a threshold of renewables penetration where grids experience a higher than acceptable likelihood of electricity reliability issues due to fluctuations in renewables availability. The task of improving the match of supply and demand in grids with deep renewables penetration is aimed directly at increasing these limits to renewables penetration.

1.1 Problem Statement

As a step toward meeting the growing challenge of limits to renewables penetration, we consider a set of methods that can be brought to bear on this problem. We first examine existing, well-studied methods for matching supply and demand, such as modifications of control algorithms used for managing existing supply resources and addition of large-scale energy storage resources. We then study the role of flexible energy loads that can actively alter their energy consumption. We look specifically at a class of flexible loads called *supply-following electricity loads*, which can change their consumption patterns to respond to variable grid conditions.

In this dissertation, we seek to answer the following questions related to supply-following electricity loads:

- To what extent can supply-following loads improve the match of supply and demand in electricity grids?
- How can we make existing loads become supply-following?
- How does the impact of supply-following loads change as renewables penetration increases?
- What are the remaining challenges for matching supply and demand on grids with large populations of supply-following loads, and how can they be addressed?

1.2 Contributions

The primary contributions of this dissertation are as follows:

- A methodology for scaling renewables penetration on a model of an electricity grid that maintains fidelity to the temporal patterns and geographic availability of existing renewables generation.
- Framing of the challenges of electricity grids with deep renewables penetration as opposed to current electricity grids, and assessment of the potential for technologies such as energy efficiency, hydroelectric management, and supply-following loads to improve operation of such grids.
- Implementations of supply-following electricity loads, including the first-ever supply-following refrigerator (to our knowledge) and an energy-efficient HVAC system that is able to sense and respond to occupancy using only temperature readings from a thermostat.
- Evaluation of large populations of multiple types of supply-following loads at the scale of a large electric grid.

- Assessment of the potential overall impact of supply-following loads to a large electric grid with deep renewables penetration.

1.3 Roadmap

In Chapter 2, we provide background on electricity grids, including an explanation of terminology related to supply and demand on electricity grids, a profile of three electricity grids under study, an anecdotal explanation of limits to renewables penetration, and an outline of metrics used for evaluation throughout the study. We continue in Chapter 3 by presenting a methodology for creating models of electricity grids with different levels of renewables penetration and using this methodology to create models of the three grids under study at deep renewables penetration. We then examine the sensitivity of these grids to assumptions made in our model, and provide a more formal definition to limits to renewables penetration. In Chapter 4, we discuss increasing the limits to renewables penetration in the context of a variety of techniques employing supply, storage, and demand resources. We continue in Chapter 5 by focusing on a specific class of demand resources called supply-following electricity loads, implementing and evaluating instances of a series of supply-following loads. We continue our evaluation in Chapter 6 by examining the potential for large populations of supply-following loads to improve the match between supply and demand on an electricity grid with deep penetration of renewables. Last, in the context of a grid with large populations of supply-following loads and deep penetration of renewables, we assess the remaining variability present in the grid, outlining an agenda for future study.

1.4 Statement of Collaborative Work

The work in this dissertation arises from collaboration with many different people, and has been published in a number of different venues:

- The data and analysis in Chapters 2 and 3 were published in IEEE SmartGridComm 2013 in collaboration with V. Smith, D. Culler, and C. Rosenberg [117].
- The data and analysis in Sections 2.2.1, 4.1, 4.2, and 4.3 were published in ACM/IEEE ICCPS 2012 in collaboration with R. Katz and D. Culler [115].
- The data and analysis in Sections 5.1, 5.2, and 6.1.1 were published in IEEE SmartGridComm 2010 in collaboration with D. Culler and P. Dutta [114].
- The data and analysis in Section 5.3 were published in the Proceedings of the IEEE Special Issue on Cyber-Physical Systems in collaboration with A. Aswani, N. Master, D. Culler, and C. Tomlin [9].
- The data and analysis in Sections 5.4.1 and 6.1.2 were published in IEEE SmartGridComm 2013 in collaboration with K. Lutz and D. Culler [116].

- The data and analysis in Section 5.4.2 were published in IEEE SmartGridComm 2012 in collaboration with N. Murthy, K. Bojanczyk, D. Auslander, and D. Culler [90].

Chapter 2

Electricity Grid Background

Modern electric grids comprise a complex distributed system in which a portfolio of electric power generation resources must be managed dynamically to meet an uncontrolled, time-varying demand. This demand must be met while adhering to constraints imposed by the transmission network, generator ramp rate capabilities, and emissions limits, with sufficient reserve to handle faults and failures. The primary control loop is typically realized through a system operator solving an iterative unit commitment problem; based on a prediction of load, an assignment of generation capacity is made on an hour-by-hour basis through a day-ahead auction that accounts for transmission limits and losses and maintains a certain amount of reserve. The matching of generation to load is refined through hour-ahead and 5-minute-ahead markets based on recently observed demand. The feedback from load to generation is ultimately manifested through observations of power quality variations, i.e., frequency fluctuations and voltage deviations resulting from any mismatch. Generators respond to mismatches by engaging more or less of the reserve. Commercially, there are widespread efforts to utilize information technology to improve the efficiency and effectiveness of the grid through so-called "smart meters" that monitor loads and report on 15-minute intervals, synchrophasors that observe power quality at intermediate points in the transmission grid, and delivery of pricing signals to trigger a response from demand. These efforts begin to introduce information planes to augment the physical planes of classic electric grids [89].

However, the integration of large amounts of fluctuating renewable resources, such as wind and solar, make the control of future electric grids fundamentally more challenging. In addition to uncertainty in demand, supplies are no longer completely dispatchable and often do not replicate the slow ramp rates and high inertia that characterize traditional generation. Some argue that to achieve deep penetration of renewable resources, the operational model of the grid must be turned around, changing the paradigm from load-following supplies to supply-following loads [74]. In contrast to the utility-centric grid that arose from the industrial revolution, this new model has been called a consumer-centric grid, with elements of networking and control more prominent. Indeed, a recent study by the California Council on Science and Technology on how to meet California's GHG emissions target of 80% below 1990 levels by 2050 concluded that the largest leap in realizing this goal is technology

to maintain the balance between load and available supply, termed "Zero-Emissions Load Balancing" [20]. The report notes that even if only the state-legislated minimum of 33% renewable energy is met, firming these fluctuating renewables with fossil fuel-based generation, often intermittent and peaker plants, would alone exceed the entire 2050 emissions allotment, notwithstanding emissions from the remainder of electricity and transportation fuels.

2.1 Terminology

In this section, we establish some terminology that is used throughout this work to describe well-understood concepts in energy systems. For a more thorough background of the electricity grid and its operation, we recommend another source [89].

2.1.1 Supply Terms

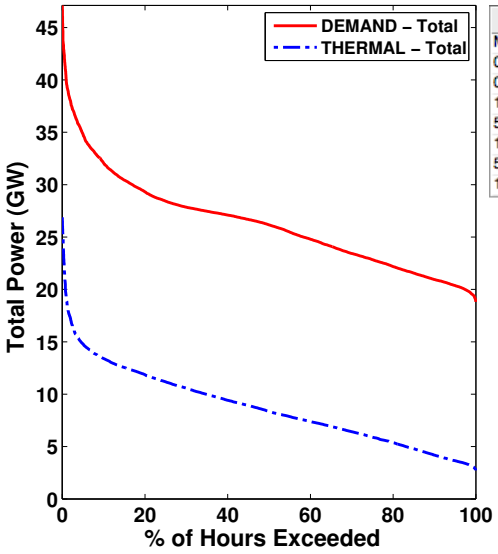
Generation facilities have traditionally been divided into three categories, based on cost of operation and agility in power generation: baseload, intermediate, and peaking. A *loading order* dictates the progression by which facilities will be turned on as additional power generation is needed. Baseload generation is generally first in the loading order, and constitutes facilities that have the lowest cost of operation but are least flexible in changing output power; these include nuclear, coal, and run-of-river hydroelectric generation. Intermediate facilities are next in the loading order and, while still operating for long periods of time like baseload generation, can alter generation on demand; examples of intermediate generation are combined-cycle gas turbines and other, generally older thermal generators. Last, peaking facilities are used only intermittently, to provide generation for short periods of time and at the highest cost of operation; they must be able to start and stop quickly. Examples of peaking facilities include gas turbines and hydroelectric facilities with reservoirs.

To manage this array of generation resources, grid operators create schedules of required supply according to predicted overall demand on the grid. Generally, to populate these schedules with contracts for generated energy from generation facilities, one or a series of markets is used. For example, in California, there are three main markets for electricity, representing sales of power at different durations prior to the provision of electricity: day-ahead, hour-ahead, and real-time or spot. The *day-ahead* market is the biggest, comprising the grand majority of total energy sales; bids represent contracts to produce a specific quanta of power for a particular hour the following day. As the demand varies from day to day, the predicted power required for an hour the following day is relatively coarse and possibly too much or too little. Since generators are given the most advanced notice of requirements relative to the other markets, the cost of energy in the day-ahead market is the cheapest among the three. Next is the *hour-ahead* market, whereby generators provide bids only one hour in advance of the need for power production. Finally, there is the *real-time* or *spot* market, which is the smallest of the three by total energy and has a bidding and

clearing process for contracts every five minutes. In addition to these markets, there are additional *ancillary services* offered by generators (and increasingly, from demand resources via *demand response*, explained below), that operate on even shorter timescales to further address discrepancies between supply and demand. Among these ancillary services are non-spinning reserve, which is capacity that must be turned on to be brought online; spinning reserve, capacity that is always available to be brought online immediately; and frequency regulation, capacity which can be added or removed from the grid in a matter of seconds. The term frequency regulation arises from the need to address disparities in supply and demand that cause deflections from the grid nominal alternating-current (AC) frequency (either 50 or 60 Hertz for nearly all electricity grids). This cascade of markets allows grid operators to successively contract generation resources as the time of generation draws nearer and the prediction of required power relative to the demand improves. As a general rule-of-thumb, the price of generation resources increases as the requirements for response time decreases (*i.e.*, fast-acting generation resources cost more).

Baseload, intermediate, and peaking resources have a range of ramp rates, representing how quickly they can be turned on and off to produce or stop producing power, and operation times, representing the duration for which the resource can produce power. Additionally, some of these resources are *dispatchable*, that is, they can be actively called into and out of operation. On the other hand, some generation facilities are *non-dispatchable* – they produce electricity when an uncontrolled resource is available, such as sunlight, wind, or tides. These non-dispatchable facilities are generally also types of renewable generation. Though other renewable sources of generation may be dispatchable – such as geothermal, biomass, biogas, and hydroelectric – non-dispatchable generation is the fastest growing generation type in the U.S. [122] Generally defined, the proportion of electricity supply that comes from renewable sources as measured by energy generated annually is called the *renewables penetration*. Pushed along by renewables portfolio standards, which dictate that electricity grids must achieve particular levels of renewables penetration by a certain year, non-dispatchable renewable generation comprises a growing component in the generation mix in many electricity grids. The emergence of non-dispatchable renewable generation sources is fundamentally changing the operation of electricity grids. Now, instead of all or nearly all of generation resources being dispatchable, the output of a larger proportion of supply resources is non-dispatchable, changing how grid operators can respond to the challenge of matching supply and demand. A potential consequence of this transformation that is unsettling to existing owners of grid infrastructure is the "stranding" of generation assets; a stranded asset is a capital investment that has become redundant before it is fully depreciated. Essentially, this is a capital investment from which less than the full value was obtained. As part of this replacement process, we also consider in our models whether a generation source is *displaceable*, or able to be supplanted by another generation source. For example, in some models, we consider baseload generation to be non-displaceable, meaning that as the proportion of renewables generation grows, it cannot replace baseload generation.

A common set of measurements for comparing the outputs of generation sources are their capacity factors, load factors, and duration curves. Capacity factor is the ratio of mean de-



%Hrs	GW	GW
Max	47.128	27.014
0.1	45.398	25.779
0.5	42.373	21.887
1.0	39.786	19.144
5.0	34.846	14.812
10.0	32.086	13.414
50.0	26.173	8.368
100.0	18.792	2.720

Figure 2.1: Duration curves for total demand and thermal generation on the California grid for a year.

livered power to rated power, and load factor is the ratio of mean delivered power to peak power. These ratios provide a summarized statistical characterization of the operation of the generation fleet. The ratios broadly reflect technological capabilities of the generation type, but are also subject to availability of fuels, outages for maintenance and repair, economic decisions, or policy directives. The duration curve is another form of statistical characterization that represents a cumulative distribution function of generation over the hours of a time period; duration curves embody the probability distribution of generation for the time period. In addition to generation of a single resource, duration curves may be used for aggregates of resources (*e.g.*, all generation or renewables generation) or various forms of consumption (*e.g.*, total consumption or commercial sector consumption). Example duration curves of the total demand and the thermal generation on the California electricity grid along with a table summarizing the values for each curve are provided in Figure 2.1. These are provided as an example – further explanation of the sources of these data are available in Section 2.2.1.

As another category of energy resources, energy storage systems are a critical component of electricity grids. Though it is important to know both power as well as energy to properly evaluate the capacity of a storage resource, storage capacity is generally quoted only as the potential to deliver power (*i.e.*, in units of $Watts$ rather than $Watt - hours$). The reason for this choice is not clear to the author.

2.1.2 Demand Terms

On the demand or load side of electricity grids, a number of different techniques have been developed to facilitate managing grid operation; this is called *demand-side management*. Energy efficiency deals with loads performing the same tasks while using less energy. Other types of demand-side management use flexibility in the demand to alter patterns of consumption; this flexibility to advance or defer energy consumption is called *slack* [114]. Demand response represents a shift of energy consumption, where energy consumption is avoided during some period and either advanced or deferred, with possible energy loss in the process. Peak shaving is a type of demand response where energy consumption is relocated from peak hours to non-peak hours, effectively reducing peak demand. Supply-following uses slack in a different way – loads alter their consumption patterns according to a supply signal such as availability of energy or price of energy. For example, when availability is high or price is low, loads consume a normal amount or more electricity, and avoid consumption when availability is low or price is high. The aggregate of many loads in supply-following operation mimics the pattern of the supply signal itself.

2.2 Electricity Grids Data

We compare electricity grids for three regions: California, Germany, and Ontario. Statistics about these three grids can be found in Table 2.1. The grids have a range of sizes and renewables penetration levels, and represent geographies with considerably different energy resources. Though the portfolios of existing generation sources differ substantially, the challenge of matching supply to highly-variable electricity demand is common to all electrical grid operators. We continue by further describing each grid, including data sources, the current state of renewables, relevant policy directives, patterns to supply and demand, and the availability of renewable resources.

2.2.1 CAISO - California

Partial deregulation of the California electricity grid in the mid-1990s required the sale of some power generation stations by the small handful of electric utilities to a number of independent power producers. Balancing authority over 80% of the electricity grid of California was granted to a single entity, the California Independent System Operator (CA ISO). CA ISO is an independent, non-profit corporation that monitors, coordinates, and controls the electrical power system in the state, managing electricity flow across the 25,865-mile transmission network and operating wholesale power markets [26]. Recently, CA ISO released hourly supply data for the ten different types of generation sources under its management starting from August, 2010 [27].

[1]Capacity factor is the ratio of mean delivered power to rated power, and load factor is the ratio of mean delivered power to peak power.

[1]The Ontario solar trace is synthesized. Further details in Section 2.2.3.

	CAISO California	**EEX Germany**	**IESO Ontario**
Mean Power	26.3 *GW*	50.8 *GW*	17.1 *GW*
Min / Max Power, Hourly	18.8 *GW* / 47.1 *GW*	24.4 *GW* / 72.6 *GW*	11.9 *GW* / 24.9 *GW*
Min / Max Power, Daily Mean	21.5 *GW* / 36.7 *GW*	32.8 *GW* / 66.2 *GW*	14.0 *GW* / 21.6 *GW*
Renewables Penetration	10.9 %	26.5 %	24.7 %
Solar Capacity	0.4 *GW*	30.8 *GW*	0.0 *GW*
Wind Capacity	2.8 *GW*	30.7 *GW*	1.5 *GW*
Capacity / Load Factor[1]*, Solar*	28.7 % / 25.3 %	10.3 % / 14.4 %	12.6 %[2] / 15.6 %[2]
Capacity / Load Factor[1]*, Wind*	29.1 % / 33.2 %	17.0 % / 21.8 %	34.8 % / 32.2 %

Table 2.1: Electricity generation statistics over one year on three electricity grids: California, Germany, and Ontario.

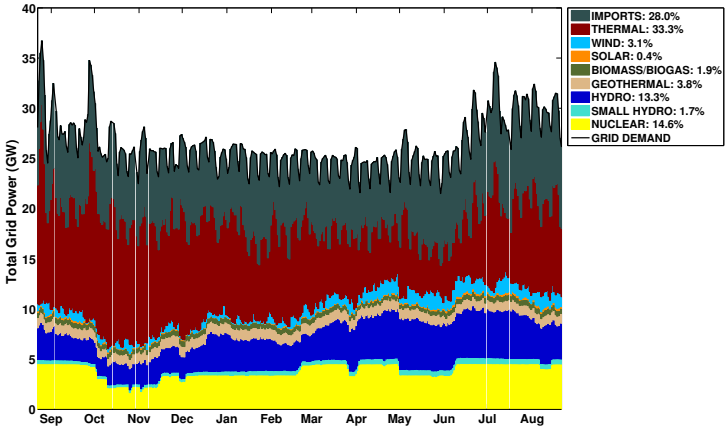


Figure 2.2: California electricity blend for a year, August 2010-August 2011. Note that biomass and biogas are combined.

Figure 2.2 shows a yearlong breakdown of the sources of electricity consumed in the CA ISO operating region, from August, 2010 to August, 2011. Overall demand, defined here as the sum of these generation sources, is also plotted; we do not consider the slight discrepancies in demand and supply that are often addressed with grid ancillary services. Electricity demand (and, thereby, production) varies on multiple timescales: *daily* with peaks in the late afternoon and nadirs in the middle of the night, *weekly* with weekends on average

Present-Day Grid			
Generation Type	**Capacity / Peak (GW)**	**Total Energy (%)**	**Capacity/Load Factors (%)**
Renewables			
Geothermal	2.600 / 1.095	3.8%	38.7% / 92.0%
Biomass/Biogas	1.145 / 0.616	1.9%	43.5% / 80.9%
Small Hydro	1.380 / 0.646	1.7%	31.7% / 67.8%
Wind	2.812 / 2.470	3.1%	29.1% / 33.2%
Solar	0.403 / 0.457	0.4%	28.7% / 25.3%
Non-Renewables			
Nuclear	4.456 / 4.581	14.6%	86.0% / 83.6%
Hydro	12.574 / 6.286	13.3%	27.7% / 55.5%
Imports	N/A / 11.055	28.0%	N/A / 66.6%
Thermal	44.339 / 27.014	33.3%	19.7% / 32.4%
Total	**80.764 / 47.128**	**100.0%**	**32.6% / 55.8%**

Table 2.2: Summary of present-day generation in California.

9.6% lower than weekdays, and *seasonally* with winter load on average 15.8% lower than air-conditioner-driven summer load.

A crucial driver to the creation of the existing blend of generation resources is the state energy policy that governs the economics of supply deployment – chiefly, California has mandated a renewables portfolio standard (RPS) target, whereby 33% of the state's generated electricity (in energy, not capacity) must come from renewable sources by the year 2020 [30]. This builds upon a previous RPS mandate of 20% of generation by 2010 [29].

Table 2.2 summarizes the characteristics of these sources over the year based on the hourly data. The capacity is based on ratings provided by the California Energy Commission (CEC) for each of the 1007 generators under CA ISO management with the exception of Imports, which reflects power purchased from other operators in the Western U.S. interconnect (WECC) [25]. Capacity factor is the ratio of mean delivered power to rated power, and load factor is the ratio of mean delivered power to peak power.

Of these generation sources, the renewables proportion consists of wind, solar, geothermal, biomass, biogas, and small hydroelectric (only facilities less than 30 MW) generation, according to the definition used by the California Energy Commission. Additionally, renewables also comprise an unpublished proportion of the imported energy, which itself comprises 28% of the total generation during the year under study. The imported energy comes from the U.S. Southwest (primarily coal and nuclear, with some solar) as well as the U.S. Pacific Northwest (mainly coal and hydroelectricity). Thermal is a mix of mostly natural gas combined-cycle and single-cycle plants providing baseload, intermediate, and peaking capacity.

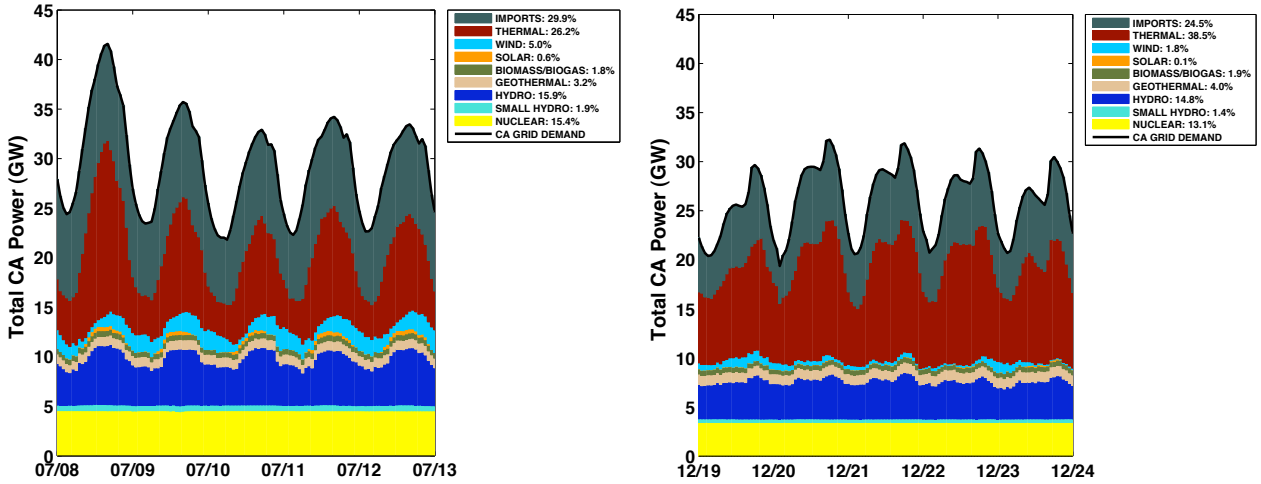
While the rule of thumb is that three times the wind or solar capacity to offset thermal generation capacity, this simple characterization demonstrates that the overall capacity factor of the thermal plants is actually lower than that of the renewables. This contrasts with common wisdom that three times as much solar or wind capacity must be deployed as compared to fossil fuel capacity. The critical difference is the degree of control over when

that capacity is dispatched.

The problem that receives the most attention in management of the California grid today is the yearly peak demand driven largely by the spike in summer afternoon air conditioning load. The capital investment to meet this summer peak is enormous. The highest 10% of demand (in excess of 42 *GW*) persists for only 41 hours a year (0.47%) and the last 20% for only 183 hours (2.1%). The generation data show that the problem of planning for load-following is even worse; an additional 33 *GW* of capacity are never called into use [25].

2.2.1.1 Temporal Variations

Several facets of Figure 2.2 reveal features of the dynamics of the grid resources. First, the nuclear generation used as a stable baseload resource in California comes from two facilities, each with two operating units. This small number of units statewide results in clearly quantized power states due to unit maintenance or repair. Second, hydroelectric sources produce more electricity in the summer, coinciding with melting of the snowpack in the Sierra Nevada mountain range. Third, much of the day-to-day variation in overall demand is met by thermal generation, which primarily includes facilities fueled by natural gas, but also some fueled by petroleum coke and coal; though many of these facilities can provide consistent baseload power, their operation suggests they are being dispatched as load-following and peaking generation to cope with demand variability. Barely visible in the figure, solar and wind, despite recently being the fastest growing sources in the generation blend, comprise only 3.5% of the total generated energy.



(a) Summer (b) Winter

Figure 2.3: Five days of present-day California electricity blend in the (a) summer and (b) winter.

Figure 2.3 shows five days of generation in summer and winter. Note the clear difference in daily peak power – the full summer and winter averages are 35.6 *GW* and 29.1 *GW*,

respectively – but also note the relative similarity of the nightly minima – the summer and winter averages are 22.6 GW and 20.2 GW. Among the resources used to cope with this variability are the aforementioned thermal generation, but also large hydroelectric generation, which can provide extra power during peak hours through scheduled generation and can be used to match supply and demand. However, we also see the characteristic patterns of the non-dispatchable renewable sources, solar and wind. Since much of this study is concerned with these generation sources, we examine them more thoroughly.

2.2.1.2 Solar and Wind Power in California

At this time, there are 403 MW of utility-operated solar capacity within the CA ISO operating region. This does not include residential or commercial rooftop solar photovoltaic (PV) deployments, of which there is currently over 1.5 GW of capacity in the state [31]. The energy generated by these panels are not reflected in supply, but instead offsets electric demand within the local distribution tree using grid-connected inverters.

Of the utility solar capacity in the CA ISO region[2], over 400 MW are from parabolic trough solar thermal facilities in southern California, with the balance coming from two solar PV sites. Additional facilities in planning and construction phases will employ advanced solar technologies, including concentrating solar thermal and PV, and panels with improved energy conversion rates.

All of the 2.8 GW of wind power farms in California are onshore, consist of low-altitude (<150m) wind turbines, and are located in 8 of California's 58 counties with roughly half the capacity in each of Northern and Southern California. Additionally, the advent of new technologies, such as high-altitude and offshore wind turbines, and the improvement of turbine efficiency will enable massive potential wind resources.

Looking more closely at the temporal variations in these non-dispatchable sources, there are important seasonal and daily patterns. Solar generation peaks during the daytime but it varies by season, as its power profile is dictated by orientation and tilt relative to Solar Normal. Its gross features are very predictable but occlusions, such as clouds, cause rapid ramps. Most of the utility solar production in California is from solar-thermal generators, so ramps are damped and generation can be delayed somewhat into the evening after the sun has gone down. Wind, on the other hand, provides power at more and varying hours of the day, but is less predictable and tends to have larger ramps in generation [114].

Figure 2.4 compares the mean total demand, solar generation, and wind generation for the hours of the day over the year. To generate each datapoint (at the hours of the day), all of the data over the full year for that hour of the day are normalized to the daily minimum; this allows for comparison of the output among all the days independent of scale, despite different seasons and other factors. This creates a distribution of normalized output at each hour of the day for each category. This figure represents the mean of that distribution. Additionally, at each data point, there is a standard deviation to that distribution that represents how

[2]Due to discrepancies in data gathered from two sources, solar peak generation exceeds rated capacity. This does not significantly alter our observations.

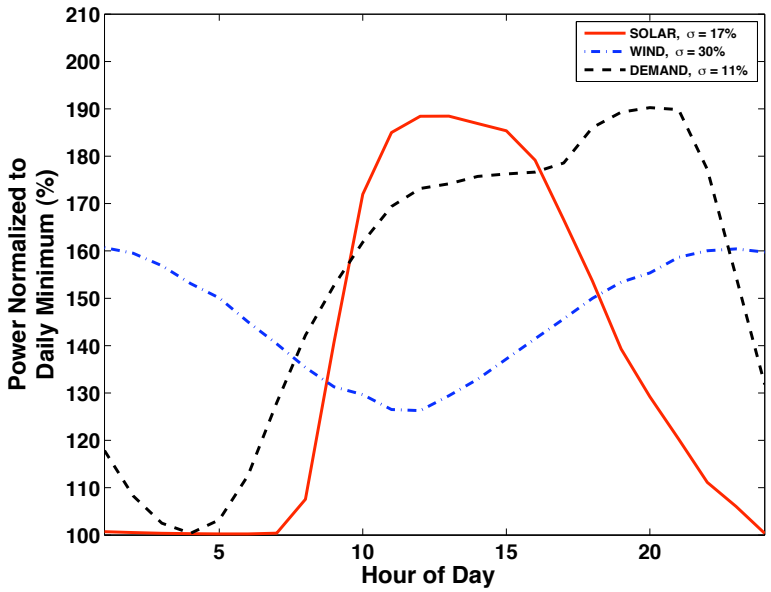


Figure 2.4: California – Mean of solar generation, wind generation, and demand (*i.e.*, the load curve) normalized to daily minimum over the hours of the day for a year. The mean standard deviation for all hours of each curve, shown in the legend, indicates relative variability.

much the generation fluctuates during that hour over the year; in the legend to the figure, the mean of those standard deviations over all of the hours of the day is provided for each category. This value represents the overall variability of that resource.

Looking at the figure, the shoulder and tail to the night in solar generation shows the impact of solar thermal generation. Wind power does tend to be higher at night and lower during the day; some have argued that the combination of wind and solar provides a good match to demand [63]. Still, the evening peak in demand represents a challenge. Furthermore, wind is less stable – the mean standard deviation across the hours of the normalized wind generation is 30%, slightly less than double the corresponding value of 17% for solar generation.

Seasonally, solar production peaks in the summer when the sun is at its most intense. Studies on the U.S. wind resource indicate that its winter production potential is double its summer potential [84]. However, the data presented in Figure 2.5 for California's wind and solar production show the opposite: far more wind energy was produced in the summer. As this analysis deals with only a year of data and is subject to the specific installations in the state, care should be exercised in drawing conclusions, nonetheless it underscores the importance of understanding the profile of local renewable resources. As these sources become a more significant proportion of the grid mixture, seasonal variability of wind and solar becomes of increasing importance.

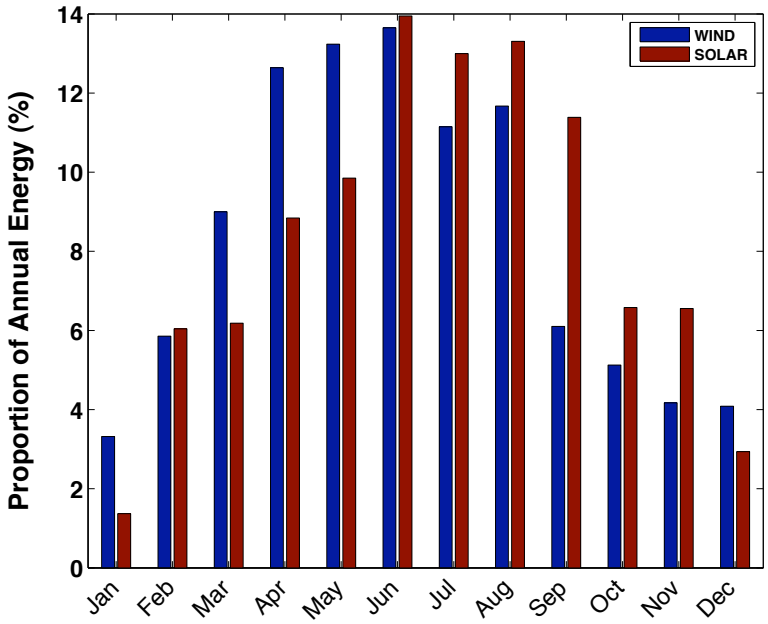


Figure 2.5: California – Monthly proportion of annual energy contribution for solar and wind generation sources.

2.2.2 EEX - Germany

The European Energy Exchange (EEX) operates the electricity market for Germany. EEX provides a Transparency Platform that publishes hourly data on generation from solar, wind, and conventional energy resources, as well as capacity data for all power generation facilities [51]. In this grid, there are generation resources with a capacity of nearly 170 GW, divided roughly in half by generators smaller than and larger than 100 MW.

Since conventional generation is not broken out into types of generation, we employ data for annual energy generated for biomass, municipal waste, hydroelectricity, and nuclear generation. Using these fractions of annual generation, we synthesize traces for these categories, assuming constant power generation for each category [3]. Though the absence of hourly data from the EEX for these types of generation results in some temporal inaccuracies, the synthesized traces provide a broad characterization of generation resources and allows estimation of the behavior of dispatchable resources in this grid. Figure 2.6 shows a year-long breakdown of the sources of electricity consumed in Germany for the year 2012. The installed renewables base of 61.5 GW, split evenly between solar and wind, is one of the largest renewables deployments in the world (second only to China). Germany has aggressively developed renewable generation sources, particularly by incentivizing a homegrown solar industry via feed-in tariffs. The German grid, which derived 16% of its energy in 2012 from nuclear generation, will see significant changes in the coming years, as recent policy decisions in response to the Fukushima disaster accelerated already-existing plans to abol-

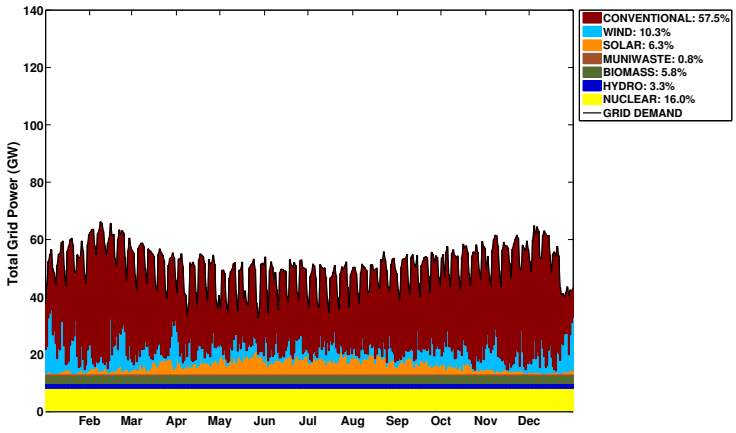


Figure 2.6: Germany electricity blend for a year, 2012.

ish nuclear power generation from the original target of 2036 to a more ambitious goal of 2022 [126].

Table 2.3 provides statistics about the operation of the different generation categories in Germany over the year. Note that for the synthesized sources – municipal waste, biomass, hydroelectric, and nuclear – the load factor has not been provided because the traces are of constant value.

Of these generation sources, the renewables proportion consists of wind, solar, biomass, municipal waste, and hydroelectric generation, a total of 26.5% of energy delivered during the year. The remaining conventional generation is primarily from lignite, hard coal, and natural gas. Germany does have 7.9 *GW* of pumped hydroelectricity storage capacity, but without operational data for these storage resources, we do not directly consider their response in our analysis. More about the opportunity for using pumped hydroelectric storage to respond to renewables availability can be found in Section 4.1.

We note a stark difference in capacity factors for wind and solar generation in comparison to the CAISO grid; the capacity factors of these sources are significantly less than those in California. Thus, to generate the same volume of energy, much more renewables capacity is required in Germany than in California, making this a relatively costlier generation category in Germany. Additionally, as Germany has already begun to phase out nuclear generation, the capacity factor for that category is less than California as well; some German nuclear assets have already been stranded.

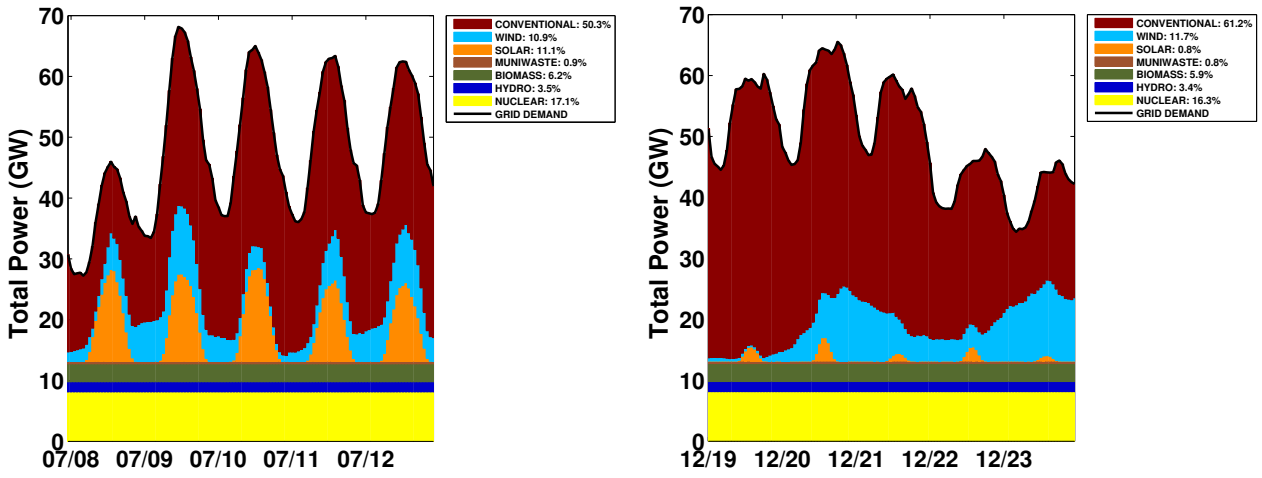
2.2.2.1 Temporal Variations

The overall demand curve in Figure 2.6 shows significant weekly variations, with large reductions in demand on the weekends, as well as large seasonal effects, with heating-driven peaks in the winter; however, these peaks are interrupted by the winter holiday season during

Present-Day Grid			
Generation Type	**Capacity / Peak (*GW*)**	**Total Energy (%)**	**Capacity / Load Factors (%)**
Renewables			
Municipal Waste	1.038 / 0.406	0.8%	39.1% / N/A
Biomass	8.920 / 2.944	5.8%	33.0% / N/A
Hydro	2.729 / 1.675	3.3%	61.4% / N/A
Wind	30.685 / 23.953	10.3%	17.0% / 21.8%
Solar	30.797 / 22.048	6.3%	10.3% / 14.4%
Non-Renewables			
Nuclear	12.078 / 8.120	16.0%	67.2% / N/A
Conventional	82.544 / 53.398	57.5%	35.4% / N/A
Total	**168.789 / 72.603**	**100.0%**	**30.1% / 69.9%**

Table 2.3: Summary of present-day generation in Germany.

which demand is noticeably depressed. Since many of the generation categories in this figure are synthesized, we cannot draw conclusions about the dispatch pattern of these assets.



(a) Summer

(b) Winter

Figure 2.7: Five days of present-day Germany electricity blend in the (a) summer and (b) winter.

Figure 2.7 shows five days of generation on the Germany grid in summer and winter. These graphs show the enormous change in overall demand from a weekend day to a week day; for example, in the summer and winter, the peak is over 20 and 15 *GW* less, respectively. Additionally, the shape of the demand is slightly different from summer to winter; the summer has smooth peaks that coincide well with available solar generation, while the winter has two peaks per day, representing the height of the workday and the rise in demand when residential consumption is at its highest in the evening; this same pattern is evident in

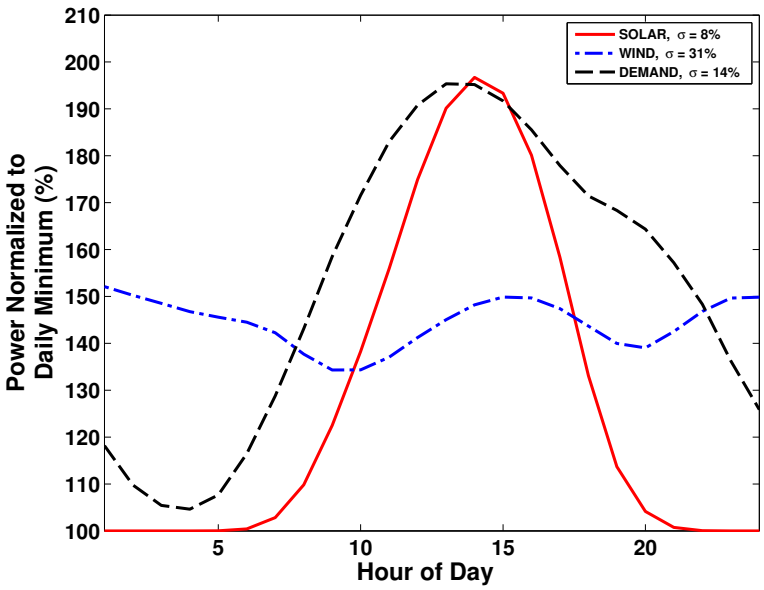


Figure 2.8: Germany – Mean of solar generation, wind generation, and demand (*i.e.*, the load curve) normalized to daily minimum over the hours of the day for a year. The mean standard deviation for all hours of each curve shows relative variability.

the other grids studied as well. We can also see the differences between the seasons in availability of the non-dispatchable renewable resources, solar and wind; we examine them more thoroughly in the following section.

2.2.2.2 Solar and Wind Power in Germany

The solar capacity of Germany is nearly 30.8 *GW* and has been doubling every 1.5 years for the last 20 years. Solar insolation is best in the southern state of Bavaria, which accordingly has the largest installed capacity of any state. A healthy feed-in tariff has produced a strong solar industry, with German companies among the most successful solar companies worldwide.

The 30.7 *GW* of wind capacity in Germany is almost entirely onshore, but of varying generations of technologies. Germany has some of the tallest and largest wind turbines in the world, and has been upgrading older technologies to newer, more efficient versions. Geographically, wind production is best in the northern states, away from the base of industrial production in the south. Recently, Germany has begun to install wind turbines offshore in the North and Baltic Seas, though these deployments comprise a tiny fraction of the overall installed base.

Figure 2.8 shows the mean total demand, solar generation, and wind generation for the hours of the day over the year. To generate each datapoint (at the hours of the day), all of the data over the full year for that hour of the day are normalized to the daily minimum; this

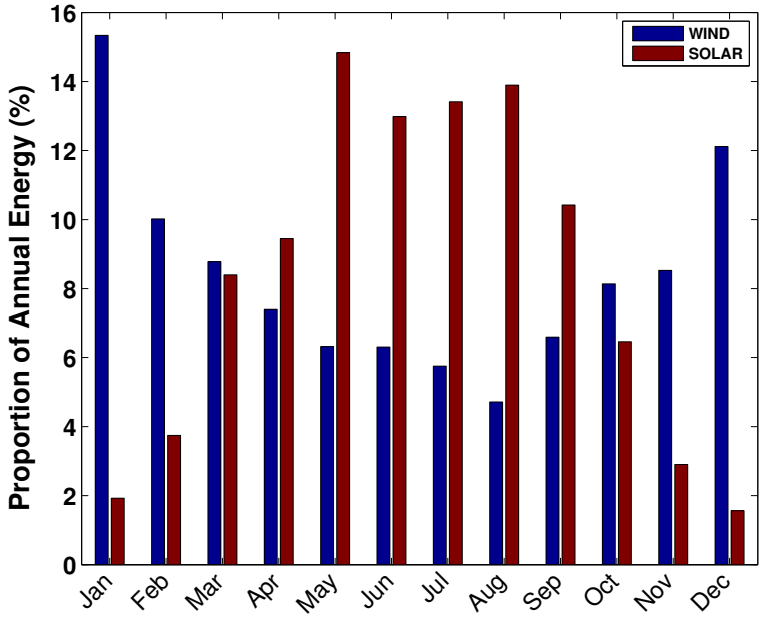


Figure 2.9: Germany – Monthly proportion of annual energy contribution for solar and wind generation sources.

allows for comparison of the output among all the days independent of scale, despite different seasons and other factors. This creates a distribution of normalized output at each hour of the day for each category. This figure represents the mean of that distribution. Additionally, at each data point, there is a standard deviation to that distribution that represents how much the generation fluctuates during that hour over the year; in the legend to the figure, the mean of those standard deviations over all of the hours of the day is provided for each category. This value represents the overall variability of that resource.

Looking at the figure, just as in California, we see that solar generation has long periods during the night hours with zero generation, whereas wind generation is often present at all hours of the day. Further, just as in California, the standard deviation in solar generation for each hour is far less than that for wind; this means that the output of solar generation is far more predictable by hour of the day. In relation to demand, the peak of solar generation coincides well with the peak in overall consumption; thus, despite a low capacity factor for solar in Germany, there is increased value of solar due to its higher affinity to the demand pattern as compared to other grids.

Seasonally. solar production peaks in the summer while wind production peaks in the winter, as shown in Figure 2.9. This opposing pattern allows these resources to be complementary on this grid, providing generation throughout the year, unlike California where seasonal availability of these non-dispatchable resources is often aligned.

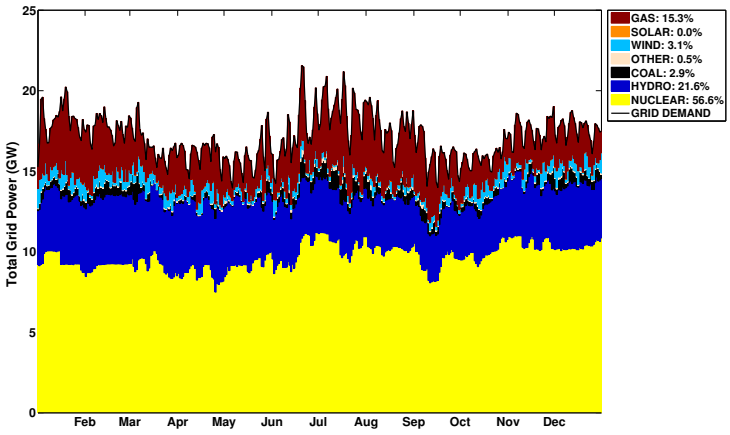


Figure 2.10: Ontario electricity blend for a year, 2012.

2.2.3 IESO - Ontario

The Independent Electricity System Operator (IESO) balances supply and demand across the electricity grid of Canada's most populous province, Ontario. The IESO provides hourly output and capability (capacity) for each of the 88 generation facilities in its territory, totaling 36 *GW* of capacity [66].

Figure 2.10 shows a yearlong breakdown of the sources of electricity consumed in Ontario for the year 2012. A majority (56.6%) of total energy is from nuclear generation, representing a very different data point from the other two grids under study. In the last decade, Ontario has taken significant steps to phase out its coal generation in favor of natural gas generation [97].

Renewables on this grid, delivering 24.7% of energy, are comprised primarily of hydroelectric generation, but also include some wind generation. As opposed to Germany, the lack of any utility-scale solar generation is interesting, given that a substantial portion of the province is further south than the southernmost point of Germany and has better solar resources.

Table 2.4 summarizes grid operation for a year in Ontario. Just as in California, the capacity factor for gas generation is less than that for wind; this represents significant excess gas capacity. Further, we can see that nuclear generation provides a large, stable baseload generation resources in Ontario. On the other hand, coal – usually a baseload generation resource as well – has a much lower capacity factor; since Ontario is aggressively phasing out coal generation, there are enormous stranded assets as many large facilities are no longer operated.

Present-Day Grid			
Generation Type	**Capacity / Peak (*GW*)**	**Total Energy (%)**	**Capacity / Load Factors (%)**
Renewables			
Hydro	7.939 / 6.029	21.6%	46.5% / 61.3 %
Wind	1.511 / 1.633	3.1%	34.8% / 32.2%
Solar	0.000 / 0.000	0.0%	0.0% / 0.0%
Non-Renewables			
Nuclear	12.998 / 11.290	56.6%	74.6% / 85.8%
Coal	3.293 / 2.377	2.9%	14.9% / 20.6%
Gas	9.987 / 6.411	15.3%	26.3% / 41.0%
Other	0.122 / 1.319	0.5%	74.3% / 6.9%
Total	**35.850 / 24.872**	**100.0%**	**47.8% / 68.8%**

Table 2.4: Summary of present-day generation in Ontario.

2.2.3.1 Temporal Variations

The annual load shape for Ontario shows both a summer, air-conditioner driven peak and a winter, heating-driven peak. The magnitudes of these peaks has been changing – until 1998, the winter peak was always higher, but since then, the summer peak has been higher for every year except one [65]. As a result of this changing demand pattern, the need for flexible resources has increased.

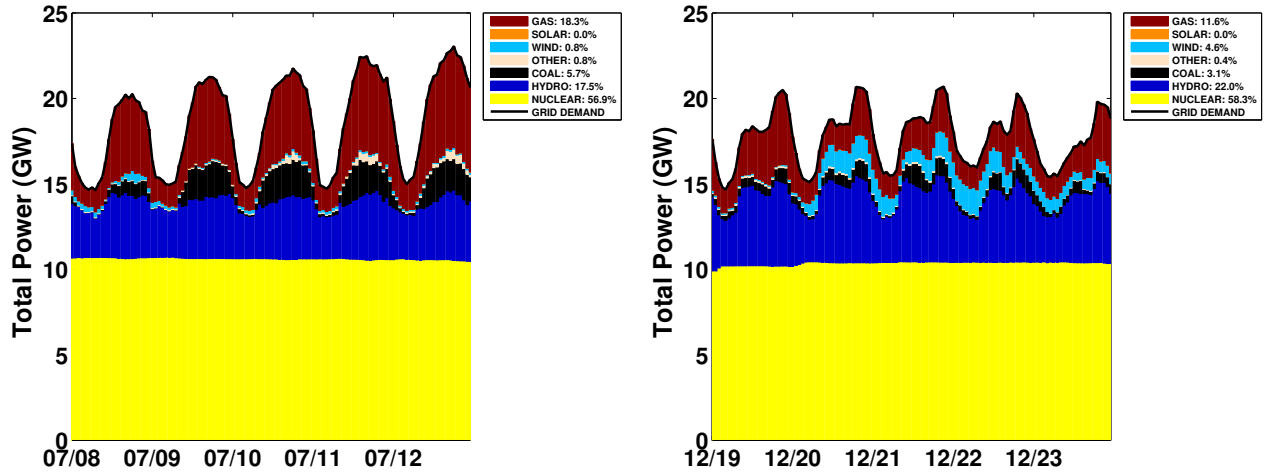
Figure 2.11 shows five days of generation on the Ontario grid in summer and winter. Like Germany, the differences in season manifest in the shape of the daily peaks; summer days have a single peak where winter days generally have two peaks. Much of the dispatchable generation is from hydroelectricity, which closely maps the demand curve. Coal is operated as an intermediate resource, turned on for the course of the day, but turned off at night; this represents a nontraditional use of coal generation. Wind generation is nearly absent in the summer, but contributes significantly in the winter.

2.2.3.2 Solar and Wind Power in Ontario

Since we desire to consider solar generation as a utility-scale generation resource, we synthesized an aggregate solar trace using generation data from PVOutput for twelve solar generation facilities located throughout Ontario [102]. This aggregation is able to provide a trace representing the geographic diversity of the solar resource available in the province, but is vulnerable to the specific characteristics of the small number of installations. In our experiments, we scale this aggregate trace to study the viability of solar generation in Ontario.

The 1.5 *GW* of wind generation in Ontario is all onshore, and is slated to grow by 200% by 2015 [97]. Wind is increasingly slated to displace coal, and will also help as older nuclear generation facilities are phased out.

Figure 2.12 shows the mean total demand, solar generation, and wind generation for the hours of the day over the year. To generate each datapoint (at the hours of the day), all of



(a) Summer (b) Winter

Figure 2.11: Five days of present-day Ontario electricity blend in the (a) summer and (b) winter.

the data over the full year for that hour of the day are normalized to the daily minimum; this allows for comparison of the output among all the days independent of scale, despite different seasons and other factors. This creates a distribution of normalized output at each hour of the day for each category. This figure represents the mean of that distribution. Additionally, at each data point, there is a standard deviation to that distribution that represents how much the generation fluctuates during that hour over the year; in the legend to the figure, the mean of those standard deviations over all of the hours of the day is provided for each category. This value represents the overall variability of that resource.

Since there is no utility-scale solar generation in Ontario, we use the synthesized solar trace in this figure. As in the other two grids, wind and solar provide very different patterns of availability. Again, solar is more suited to demand than wind, but in this case, solar peaks slightly earlier than the first peak in demand, and the second peak in demand occurs well after solar generation is available. Just as in the other grids, the standard deviations of the wind resource are far more than the solar resource.

Looking at wind and solar generation in Ontario over the year in Figure 2.13, we see a similar pattern to Germany, where solar peaks during the summer, but wind peaks during the winter. Again, this represents an opportunity for these resources to complement each other.

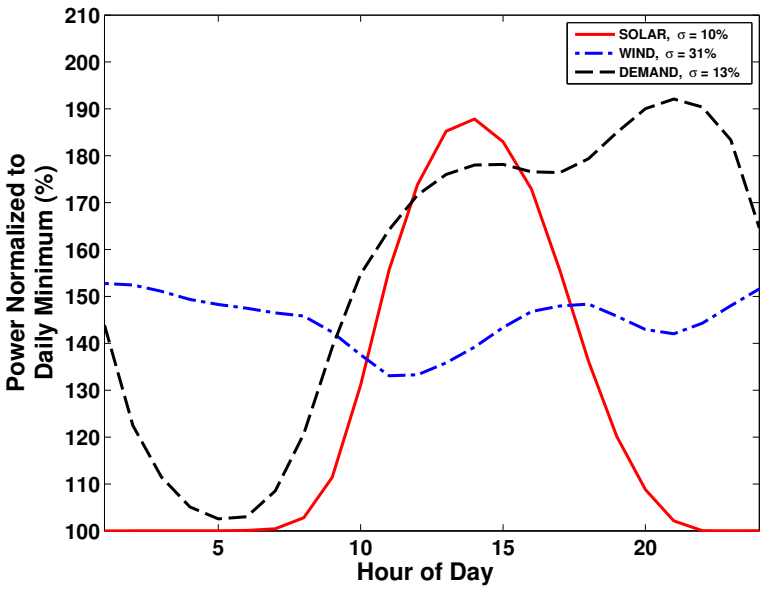


Figure 2.12: Ontario – Mean of solar generation, wind generation, and demand (*i.e.*, the load curve) normalized to daily minimum over the hours of the day for a year. The mean standard deviation for all hours of each curve shows relative variability.

2.3 Limits to Renewables Penetration

As the supply mix for electricity grids changes from one comprised almost entirely of dispatchable, fossil-fuel powered sources to one balanced between those sources and non-dispatchable, renewable sources, grids are becoming more difficult to operate, causing unexpected consequences. These consequences are a result of *limits to renewables penetration*. We introduce this concept through two anecdotal examples before formalizing the description in Section 3.6.

In 2012, the German grid, which had its highest solar capacity ever, at 30.8 GW, yielded higher emissions of greenhouses gases (GHGs) than the previous year [100]. The renewables-friendly policies adopted by the German government dictate that solar and wind receive first priority to supply electricity. However, the high rates of solar generation during the day caused pumped hydroelectric storage facilities, long profitable in the country, to operate at a fraction of their previous capacity, as electricity was no longer needed during the day. With pumped storage facilities no longer providing firming to the grid and nuclear energy generation reduced after the Fukushima disaster, the need for stable sources of power increased. In response, the owners of conventional generation resorted to using more of the lowest-cost generation, which in Germany is brown coal or lignite. Some of these facilities had even been scheduled to shut down permanently, but were called back into operation. These sources of energy tend to have among the highest production of GHGs, resulting in a net overall increase in GHG emissions. Further increasing the capacity of renewables on the German

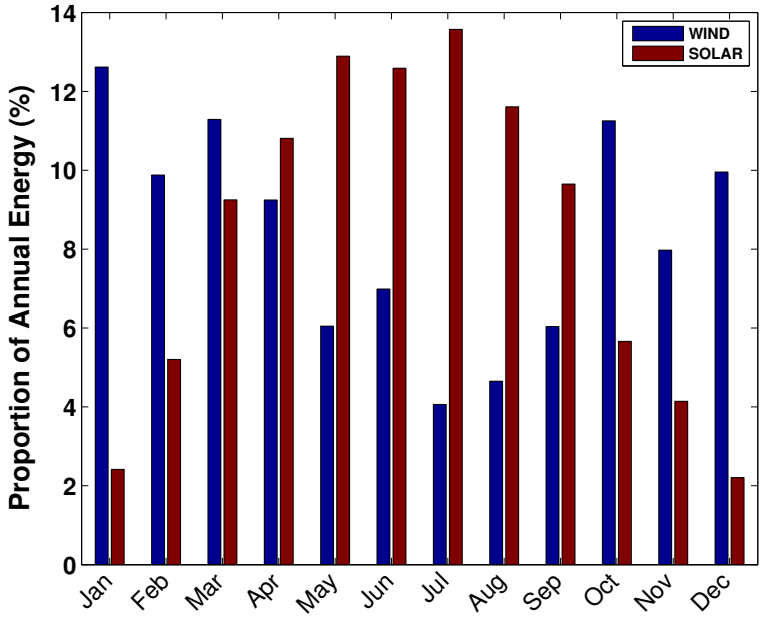


Figure 2.13: Ontario – Monthly proportion of annual energy contribution for solar and wind generation sources.

grid will only exacerbate this problem. Already, German electricity consumers are suffering under the burden of costs related to increasing renewables penetration to the existing level, with the highest electricity prices in Western Europe [45]. This leads to a principle describing limits to renewables penetration: at some level of renewables penetration, the benefits of increased renewables capacity begin to diminish, and eventually the cost becomes too expensive to bear.

As another instructive example, the Danish electricity grid in 2007 had substantial wind capacity of roughly 3.1 GW, providing roughly 20% of consumed electricity [50]. However, the Danish grid was and still is small, with an average demand around 4 GW. To deal with large fluctuations in wind generation, grid operators buy and sell power from its European neighbors in times of deficit and surplus; generally, this means expensive purchase of German and Nordic hydroelectric power and cheap sales of wind generation to the same neighbors' pumped hydroelectric storage facilities. However, this arrangement only works with the unique geography of this region, with multiple neighbors possessing enormous hydroelectric resources, and the relatively small Danish grid [88]. In most cases, the best sites for pumped hydroelectric storage facilities are already developed, so capacity of this type is not expected to increase substantially. Thus, this plan is not a generally-applicable blueprint for attaining high levels of renewables penetration. This leads to a second principle of limits to renewables penetration: limits to renewables penetration are highly grid-specific, but as renewables penetration increases, the measures required to maintain a sustained match between supply and demand on a grid become more draconian.

2.4 Metrics for Evaluation

To address the challenge of increasing renewables penetration at bearable levels of cost, this thesis evaluates a suite of techniques for increasing limits to renewables penetration. Among these, we consider changes to operation of existing generation sources, provisions of vast reservoirs of energy storage, and additions of a promising class of demand-side resources called supply-following electricity loads.

To evaluate the efficacy of these techniques for increasing limits to renewables penetration, we use two different metrics: decrease in natural gas generation while maintaining a given level of renewables penetration and decrease in renewables capacity while maintaining a given level of renewables penetration. In the former case, any gains from incorporating a change in load behavior reduce the need for natural gas facilities to operate, and in some cases may obviate the need for some facilities to exist. In the latter case, the original capacity of renewables required to maintain a level of renewables penetration can be reduced. The differences between these two metrics are a result of the affinity of each timeseries of renewables generation with the timeseries of total demand.

In looking at natural gas generation, we observe the total generation as well as the duration curve of generation. However, the duration curve is for aggregate generation of the entire natural gas generation fleet; we leave it to future work to arrive at unit commitment schedules of natural gas generation facilities.

Chapter 3

Methodology for Modeling Grids with High Renewables Penetration

As a vehicle for evaluating techniques for increasing limits to renewables penetration, we require a method to model the operation of an electricity grid at varying levels of renewables penetration. This methodology has three critical requirements:

- It must maintain fidelity to the grid context. The supply mix chosen for a grid reflects existing technological, economic, and policy conditions in the context of the grid. Further, the choice of renewables technologies is affected by availability of those resources in a specific geography. These characteristics must be taken into consideration in scaling up the renewables penetration of a grid.

- It must produce a timeseries representation, as opposed to a statistical model. Limits to renewables penetration are often governed by performance during extreme, outlier events – in fact, power capacity is nearly always sized to peak demand, the leftmost extreme of the duration curve as seen in Figure 2.1. Statistical models, often used in policy-oriented studies of electrical grids, aim to capture outlier behavior by attaching probability to those events and either running simulations or operating on distributions. However, in practice, outlier events concerning peak demand on electricity grids often occur in succession. For example, in air-conditioning-driven grids like Ontario and California, peak demand is nearly always correlated with the hottest days of the year, and occurs during the few early evening hours of those days. A stretch of outlier events is difficult to capture in a statistical model, but much more prevalent in a temporal model. It is precisely these outlier events, spanning multiple hours, that embody the limits to renewables penetration. Understanding these events is necessary to develop a viable control regime with deep penetration of fluctuating renewables.

- It must be consistent among different grids. Comparison among studies modeling electrical grids is often challenging because of varying assumptions taken in analysis. For comparison of the effectiveness of different strategies for increasing limits to renewables

penetration, the assumptions behind grid models should be identical. By employing different assumptions, such as the knobs available to the model or the scaling behavior of specific renewable resources, comparison of results of the models loses its meaning. Therefore, we must maintain the same assumptions to compare the effectiveness of various strategies for increasing limits to renewables penetration on different grids.

Given these requirements, this chapter describes a methodology for scaling renewables penetration that is consistent between grids, maintains the geographic characteristics that govern renewables generation in each grid, and produces a representation that captures temporal dynamics. We then examine the ramifications of this methodology for balancing between solar and wind generation in each of the three grids under study. We employ this analysis to scale each grid to a deep level of renewables penetration, 60%, and analyze the resulting supply models. We continue by relaxing some of the assumptions taken in attaining a balance between solar and wind generation and choosing non-displaceable categories of generation. We then formalize the notion of limits to renewables penetration, producing definitions for hard and soft limits, before surveying related techniques.

3.1 Scaling Methodology

In this section we present a methodology for estimating the effects of altering renewables penetration on electricity grids using temporal generation data. We make three key initial assumptions: we are given timeseries and capacity of each generation type; the demand curve is fixed and unresponsive to availability or cost of electricity (*i.e.*, no demand response), and thus, aggregate supply equals demand; and renewables generation can be scaled proportionally to increase generation. Given these assumptions, we seek to understand how to meet the demand curve by using mixes of energy sources other than the original one.

We treat time as discrete, indexed by time slot i. For each grid we are given a yearly supply curve $\mathfrak{S} = (S_1, \ldots, S_K)$, *i.e.*, a timeseries of what energy was supplied in the grid on an hourly basis. Let $E_d = \sum_{i=1}^{K} S_i$ be the total energy delivered in the year where K is the number of hours in a year. We also know the original mix of sources, *i.e.* $S_i = \sum_{j=1}^{J} s_{i,j}$ where the grid has access to J sources of energy and $s_{i,j}$ is the amount of energy supplied by source j in time slot i.

With fixed demand and no major demand response, we assume the supply curve $\mathfrak{S}$ equals the demand curve $\mathfrak{D} = (D_1, \ldots, D_K)$, ignoring ancillary services. To model increasing penetration, we treat the regulator as imposing a penetration rate of renewables of p, denoting that the portion of total energy *delivered* to consumers by renewable sources over the year should be pE_d. Though we scale just two types of non-dispatchable renewable sources, solar and wind, more types could be considered if needed. For each of these two sources, we know the existing capacity in the grid under consideration, say $\mathfrak{C}_\mathfrak{s}$ for solar and $\mathfrak{C}_\mathfrak{w}$ for wind and the timeseries $\mathfrak{E}_\mathfrak{w} = (e_w(1), \ldots, e_w(K))$ and $\mathfrak{E}_\mathfrak{s} = (e_s(1), \ldots, e_s(K))$ where $e_w(i)$ is the energy generated by the wind in time slot i and $e_s(i)$ is the energy generated by solar in time slot

i. By selecting a factor $a \geq 0$, we can proportionally scale the existing capacity of either of these sources, resulting in a timeseries that reflects the geographic diversity of the resource in the current grid.

In our methodology with fixed $\mathfrak{D}$, increased renewables generation *displaces* some existing source of energy that may be too polluting, expensive, or dangerous to operate. We group the original mix of energy sources into 3 categories, renewables $\mathfrak{R}$, baseload generation $\mathfrak{B}$, *i.e.*, the set of sources of energy that cannot be replaced easily, and displaceable generation $\mathfrak{U}$, *i.e.*, the set of sources of energy that can be replaced. Hence, we can now rewrite S_i as $r_i + b_i + u_i$.

We now consider a mix of renewables $(a\mathfrak{C}_\mathfrak{s}, b\mathfrak{C}_\mathfrak{w})$. This mix generates in each time slot i a total energy equal to $ae_s(i) + be_w(i)$. In the present-day mix, $a = b = 1$ and the mix is thus $(\mathfrak{C}_\mathfrak{s}, \mathfrak{C}_\mathfrak{w})$. We use this energy in this time slot to displace (if possible) u_i. Hence, if $ae_s(i) + be_w(i) - r_i \geq u_i$, we can displace u_i completely and create an excess of renewables in this time slot corresponding to $x_{a,b}(i) = ae_s(i) + be_w(i) - r_i - u_i$. If $ae_s(i) + be_w(i) - r_i < u_i$, then some of the original displaceable energy must be maintained, *i.e.*, $u_i - (ae_s(i) + be_w(i) - r_i)$. Let $Y(a,b)$ be the total amount of displaced energy, *i.e.*, $Y(a,b) = \sum_{i=1}^{K} \min\{u_i, ae_s(i) + be_w(i) - r_i\}$ and $X(a,b)$ be the total amount of excess energy, *i.e.*, $X(a,b) = \sum_{i=1}^{K} (-u_i + (ae_s(i) + be_w(i) - r_i)^+$

A mix of renewables $(a\mathfrak{C}_\mathfrak{s}, b\mathfrak{C}_\mathfrak{w})$ is **p-feasible** $(0 \leq p \leq 1)$ if $Y(a,b) + \sum_{i=1}^{K} r_i \geq pE_d$ $(a \geq 0$ and $b \geq 0)$. Note that p cannot take any values. At best, we can displace the whole existing displaceable energy $E_u = \sum_{i=1}^{K} u_i$. Let $E_r = \sum_{i=1}^{K} r_i$ be the whole existing (in the original mix) renewable energy. Then we can define a *hard limit* to renewables penetration: $p \leq \frac{E_r + E_u}{E_d}$. We discuss this concept further in Section 3.6.

A natural question in this framework is choosing the generation mix among all the p-feasible mixes of renewables that minimizes total cost. This allows us to reduce the number of dimensions of the problem to one. We assume that the cost of a renewables category is proportional to its capacity, *i.e.*, the total cost of solar capacity $a\mathfrak{C}_\mathfrak{s}$ (resp. $b\mathfrak{C}_\mathfrak{w}$) is $\pi_s a\mathfrak{C}_\mathfrak{s}$ (resp. $\pi_w a\mathfrak{C}_\mathfrak{w}$) where π_s and π_w are known positive numbers. Let the cost obtained through this minimization be Γ_{min}. We are also interested in observing the total excess $X(a,b)$ for a given target cost $\Gamma \geq \Gamma_{min}$.

Hence we can write the problem as: given $\mathfrak{S}$ and the original mix, given $p \leq \frac{E_r + E_u}{E_d}$, given $\mathfrak{E}_\mathfrak{s}$ and $\mathfrak{E}_\mathfrak{w}$, given the set $\mathfrak{F}$ of p-feasible renewables mixes $m(a,b)$, solve:

$$\mathfrak{P}_1 \qquad \min_{m(a,b) \in \mathfrak{F}} \pi_s a\mathfrak{C}_\mathfrak{s} + \pi_w b\mathfrak{C}_\mathfrak{w} \tag{3.1}$$

Let Γ_{min} be the solution of $\mathfrak{P}_1$ and the renewables capacity corresponding to Γ_{min} be $\Delta^*(p) = a^*\mathfrak{C}_\mathfrak{s} + b^*\mathfrak{C}_\mathfrak{w}$ where a^* and b^* are the solutions of the problem. Note that $\Delta^*(p)$ is a function of $\frac{\pi_w}{\pi_s} \triangleq \Theta$.

3.2 Renewables Mix and Penetration

To better understand the concept of p-feasibility, we consider $\mathfrak{F}$, the set of p-feasible renewables mixes. This set represents all possible mixes $(a\mathfrak{C}_\mathfrak{s}, b\mathfrak{C}_\mathfrak{w})$ that meet our criteria; here,

we seek to find the minimum b that results in a p-feasible mix given a. The results of this exercise on each of the three grids under study for four different values of p, representing renewables penetrations of 40%, 50%, 60%, and 70%, are in Figures 3.1, 3.2, and 3.3.

There is a key assumption that underlies these figures: baseload generation $\mathfrak{B}$ is zero. This means that in our methodology, renewables can displace any type of generation, simply following the chosen unloading order. This results in fairly evenly spaced parametric curves in these figures representing different penetration levels p. We examine the results of relaxing this assumption in Section 3.5.

On each chart, using an assumption of cost parity of renewables resources, the minimum total renewables capacity $\Delta^*(p)$ is marked with an asterisk. The locations of these minima indicate the relative strength of each type of generation given its magnitude in the geographic location and its affinity for the demand curve of the grid. For example, in the case of California, the minimum capacity mixes consist of both solar and wind, as these resources have roughly equivalent capacity factors and can address substantial yet different proportions of electricity demand. However, in both Germany and Ontario, the least-capacity mixes consist entirely of wind generation at all penetration levels; this is because the solar generation in these northerly locales has a very low capacity factor, and at each increment it is more fruitful to add wind generation. Further, the time limitations of solar generation, *i.e.*, only when the sun is up, further marginalize its effectiveness in meeting demand. Notice that at higher levels of penetration, mixes composed entirely of solar generation are not p-feasible. We explore this limit p^*, the maximum penetration where a solar-only renewables fleet can meet p, in Section 3.6.

Cost parity is another important assumption. This indicates that the cost π_s of adding incremental solar generation is equal to the cost π_w of adding incremental wind generation. In practice, this may not be the case; technological, economic, and policy considerations may contribute to a fluctuating cost profile for specific technologies, resulting in differing values for π_s and π_w over time. We explore the effects of varying these parameters in Section 3.4.

Another crucial assumption is that the cost of electricity from different sources is equivalent. Thus, displacing power from a natural gas, coal, or nuclear facility results in the same reduction in overall costs. This is certainly not true in practice, but in the interest of targeting specific categories of generation, it is possible to change the unloading order accordingly. Also, by not considering ramping requirements of different types of generation, it may make the generation schedules produced infeasible. To fully show this, it would be necessary to understand the specific limitations of each facility on the grid and see if a viable schedule can be produced. We leave this to future work.

3.3 Characterization of Grids with 60% Renewables

With a description of our scaling methodology and an understanding of the range of capacity mixes that can be used, we can now create scaled models of the grids under study at different levels of renewables penetration that combine wind and solar generation. Fig-

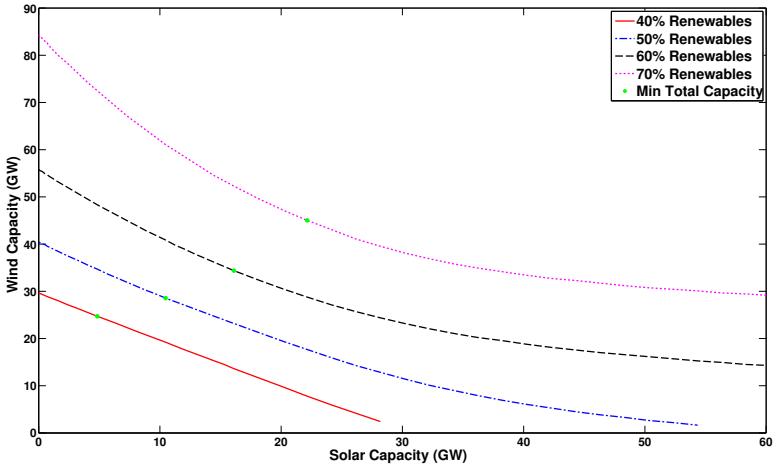


Figure 3.1: CAISO - California. Mix between wind and solar capacity at four renewables penetration levels. Note that minimum total capacity here assumes equal cost of incremental solar and wind generation. Given this assumption, the optimal generation mix at all levels of renewables penetration contains both solar and wind generation.

ures 3.4, 3.5, and 3.6 show the temporal dynamics of a "scaled" grid supply blend over a year for the three grids under study following the methodology above, with renewables scaled to 60% and a solar-wind mixture that minimizes the overall renewables capacity according to Figures 3.1, 3.2, and 3.3. The resulting capacity and load factors are summarized in Tables 3.1, 3.2, and 3.3.

Scaled Scenario - 60% Renewables			
Generation Type	**Capacity / Peak (*GW*)**	**Total Energy (%)**	**Capacity/Load Factors (%)**
Renewables			
Geothermal	2.600 / 1.095	3.8%	38.7% / 92.0%
Biomass/Biogas	1.145 / 0.616	1.9%	43.5% / 80.9%
Small Hydro	1.380 / 0.646	1.7%	31.7% / 67.8%
Wind	34.379 / 26.483	36.0%	27.6% / 35.8%
Solar	16.104 / 18.280	16.7%	27.2% / 24.0%
Non-Renewables			
Nuclear	4.456 / 4.372	1.1%	6.3% / 6.4%
Hydro	12.574 / 6.192	9.7%	20.4% / 41.4%
Imports	N/A / 10.433	7.3%	N/A / 18.5%
Thermal	44.339 / 26.677	21.8%	13.0% / 21.5%
Total	**116.978 / 47.128**	**100.0%**	**22.5% / 55.8%**

Table 3.1: Summary of generation in California, scaled to 60% renewables.

Looking at the resulting electricity breakdown for California over the year, we see the large effect of seasonal variations on the availability of renewables. The critical constraint is no longer summer peak demand, but rather the winter lulls in renewable generation. For

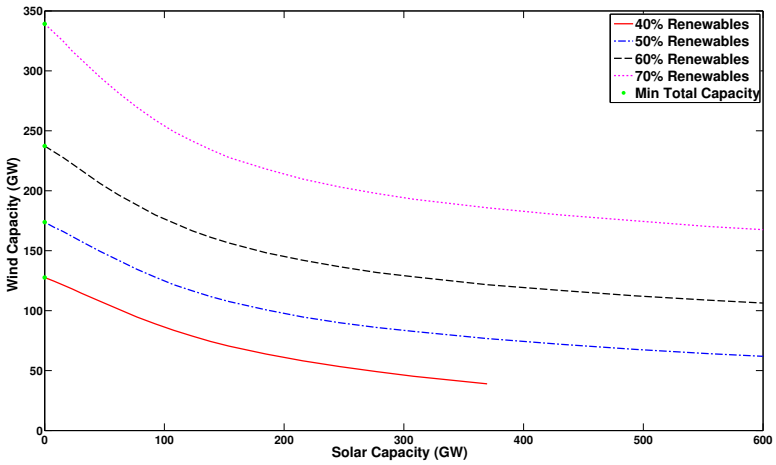


Figure 3.2: EEX - Germany. Mix between wind and solar capacity at four renewables penetration levels. Note that minimum total capacity here assumes equal cost of incremental solar and wind generation. Given this assumption, the optimal generation mix at all levels of renewables penetration contains only wind generation.

Scaled Scenario - 60% Renewables			
Generation Type	**Capacity / Peak (*GW*)**	**Total Energy (%)**	**Capacity/Load Factors (%)**
Renewables			
Municipal Waste	1.038 / 0.406	0.8%	39.1% / N/A
Biomass	8.920 / 2.944	5.8%	33.0% / N/A
Hydro	2.729 / 1.675	3.3%	61.4% / N/A
Wind	166.008 / 66.187	47.1%	14.4% / 36.1%
Solar	15.398 / 11.024	3.0%	10.0% / 14.0%
Non-Renewables			
Nuclear	12.078 / 8.120	2.5%	10.5% / N/A
Conventional	82.544 / 53.398	37.5%	23.0% / N/A
Total	**288.714 / 72.603**	**100.0%**	**17.6% / 69.9%**

Table 3.2: Summary of generation in Germany, scaled to 60% renewables.

part of the year, energy is produced in excess of demand. The new challenge in summer is to utilize the surplus of renewable energy, which could be sloughed, exported, or used to enable new energy-agile practices in industries using intermittent, seasonal, inexpensive electricity. As the penetration of solar and wind increases beyond 60%, both the existence and the magnitude of this surplus energy will increase. The detail on five days of summer in Figure 3.7(a) shows intra-day discrepancies between renewables generation and the present-day demand curve, as there is excess production for parts of the day, generally correlated with periods of solar generation, with occasional occurrences of fossil fuel-based generation needed to meet demand. On the other hand, there are months of winter where thermal and even imported generation are necessary to meet demand, such as in Figure 3.7(b). This is a result of irregular and highly seasonal solar availability and high variance in wind production

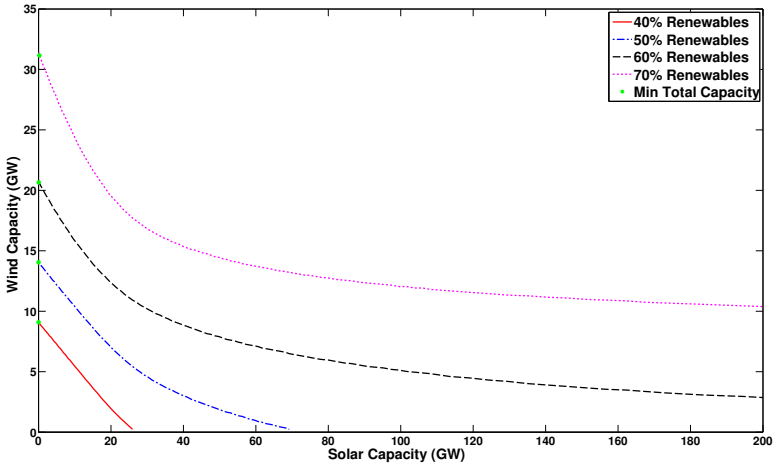


Figure 3.3: IESO - Ontario. Mix between wind and solar capacity at four renewables penetration levels. Note that minimum total capacity here assumes equal cost of incremental solar and wind generation. Given this assumption, the optimal generation mix at all levels of renewables penetration contains only wind generation.

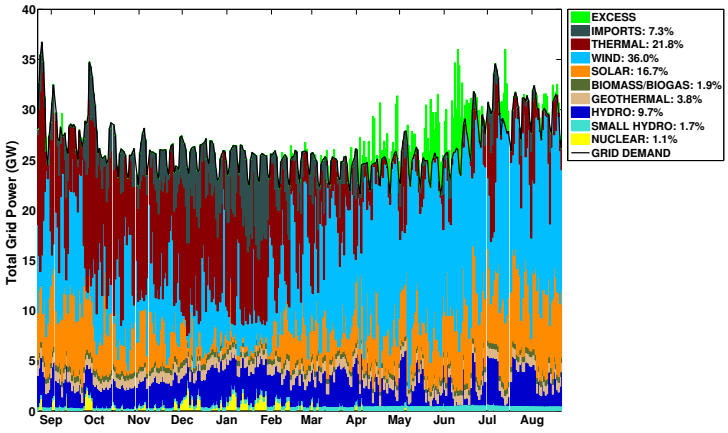


Figure 3.4: California electricity blend for a year, August 2010-August 2011, scaled to 60% renewables penetration with no baseload generation ($\mathfrak{B} = 0$). Note that biomass and biogas are combined.

coupled with reduced hydroelectric resources. Chapter 2 provides further analysis of the effects this seasonal pattern has on generation from renewables and, as a result, fossil fuels.

The yearly grid operation in Germany follows a similar maxim as California – peak renewables production coincides with peak demand; however, in the case of Germany both occur in winter rather than summer. Using the wind and solar mix with the minimum total renewables capacity from Figure 3.2, all of the renewables in the scaled Germany grid are

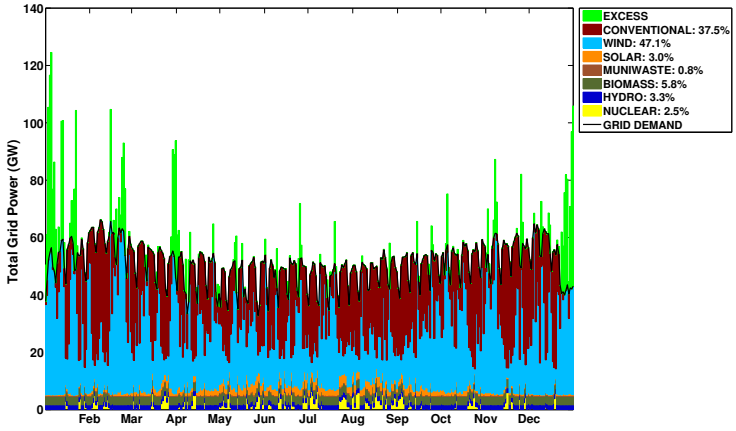


Figure 3.5: Germany electricity blend for a year, 2012, scaled to 60% renewables penetration with no baseload generation ($\mathfrak{B} = 0$).

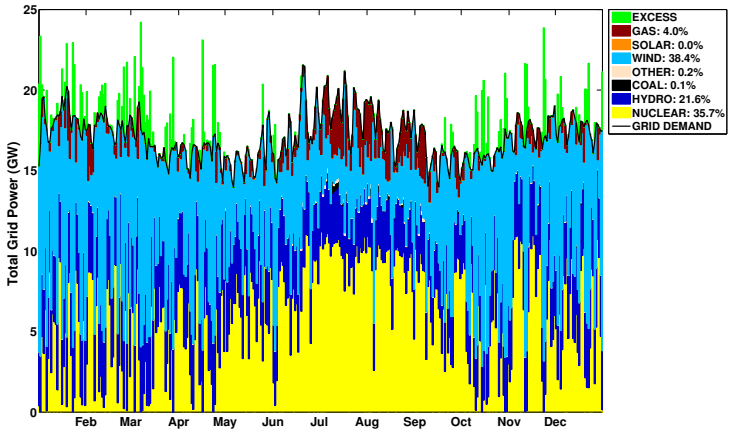
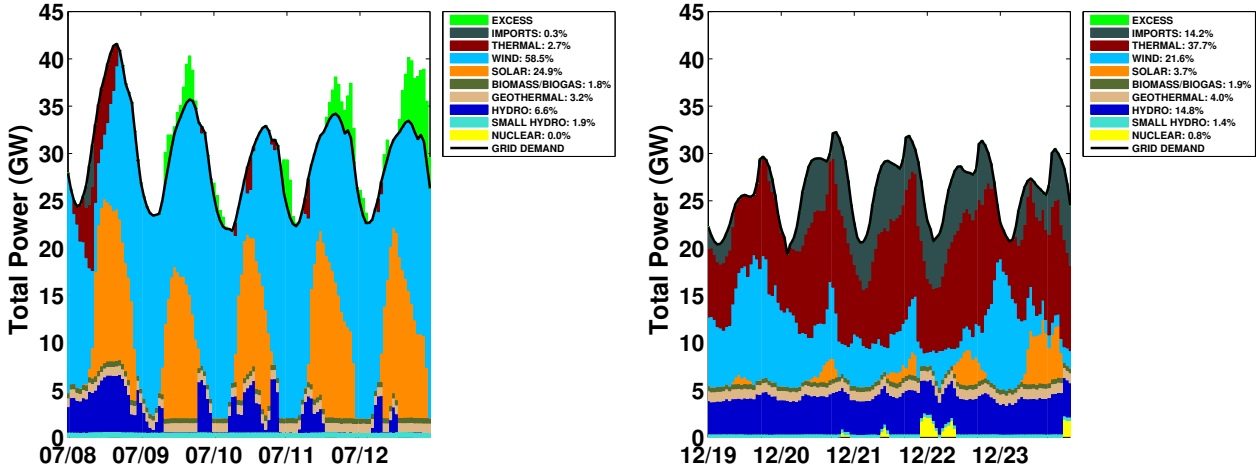


Figure 3.6: Ontario electricity blend for a year, 2012, scaled to 60% renewables penetration with no baseload generation ($\mathfrak{B} = 0$).

wind generation, which peaks in the winter (see Figure 2.9) and produces electricity at all times of day. Looking at Figure 3.8(a), which shows five days of operation during the summer in the Germany grid scaled to 60% renewables annually, we see only intermittent periods where renewables are able to meet the full demand, with most hours requiring conventional generation (primarily coal) and some hours using nuclear generation as well. Since this grid's renewables energy is almost entirely from wind, the need for conventional generation is almost entirely driven by variations in wind generation. Looking at five days in winter in this same scaled 60% renewable grid in Figure 3.8(b), we see the same pattern of intermittent

Scaled Scenario - 60% Renewables			
Generation Type	**Capacity / Peak (*GW*)**	**Total Energy (%)**	**Capacity/Load Factors (%)**
Renewables			
Hydro	7.939 / 6.029	21.6%	46.5% / 61.3 %
Wind	20.658 / 17.623	38.4%	31.8% / 37.3%
Solar	0.054 / 0.044	0.0%	12.4% / 15.3%
Non-Renewables			
Nuclear	12.998 / 11.277	35.7%	43.0% / 54.1%
Coal	3.293 / 1.897	0.1%	0.7% / 1.2%
Gas	9.987 / 6.142	4.0%	6.8% / 11.1%
Other	0.122 / 1.319	0.2%	29.7% / 2.7%
Total	**55.051 / 24.872**	**100.0%**	**31.1% / 68.8%**

Table 3.3: Summary of generation in Ontario, scaled to 60% renewables.

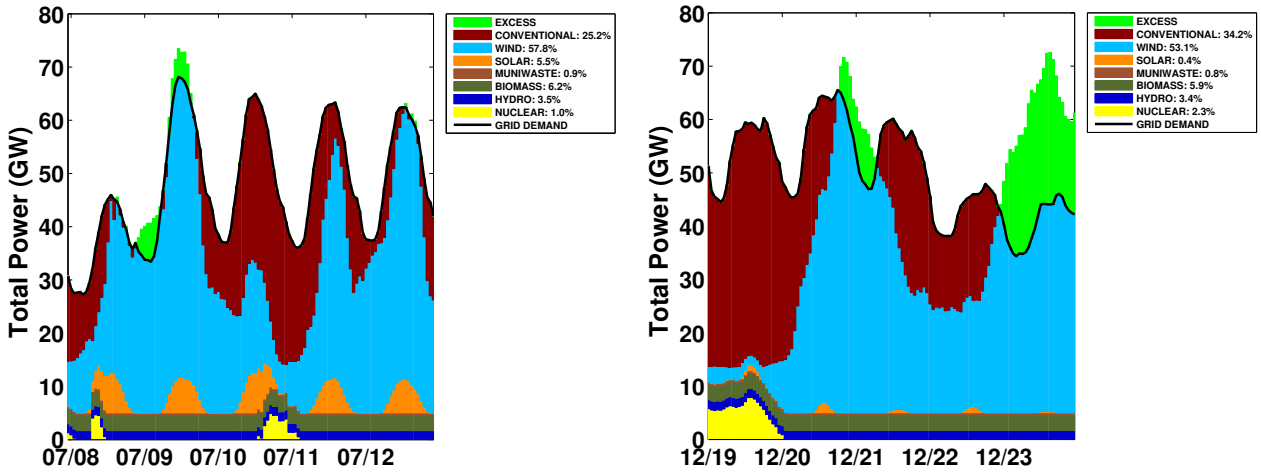


(a) Summer (b) Winter

Figure 3.7: Five days of California electricity blend scaled to 60% renewables annually in the (a) summer and (b) winter. Note that the grid does not maintain this level of renewables penetration throughout the year.

deficit and surplus, though the surplus is more drastic and virtually all renewable energy is from wind, as solar has minimal generation in winter.

The scaled Ontario grid has a lot in common with the Germany grid, in that the selection of renewable mix is entirely wind and in that locale, wind peaks in the winter (as seen in Figure 2.13). However, demand in Ontario has peaks both in the summer and the winter; the summer peak is larger. We see a similar pattern of intermittent excess and deficit, with the deficit most pronounced in the summer months. Still, the higher capacity factor of wind instead of solar is enough to counteract the seasonal strength that solar has in



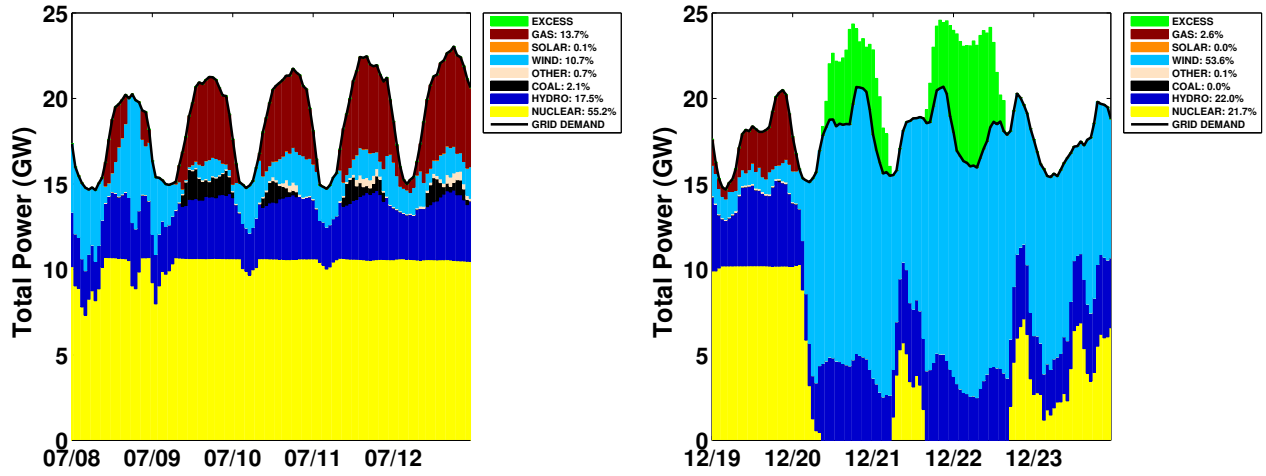
(a) Summer (b) Winter

Figure 3.8: Five days of Germany electricity blend scaled to 60% renewables annually in the (a) summer and (b) winter. Note that the grid does not maintain this level of renewables penetration throughout the year.

addressing summer shortfalls in energy. Looking more closely at five days in the summer in Ontario in Figure 3.9(a), we can see the dominance of nuclear generation, providing more than half of generation for nearly all hours. The fluctuations in wind generation necessitate the use of Ontario's dwindling coal fleet, and natural gas generation provides the dispatchable generation needed. Looking at five days in winter in Figure 3.9(b), we see a massive increase in the availability of wind generation, displacing the need for natural gas and nuclear generation during most hours. However, the fluctuations in this wind source still require significant natural gas capacity to respond to periodic shortfalls in wind generation.

Common to all of the scaled grid models so far is the notion that there is no baseload generation, *i.e.*, any type of generation can be displaced when adding renewables. With this assumption, certain elements of the portfolio may be operated outside their technological limitations. For example, a nuclear facility generally has a very slow ramp up, taking multiple days to start from off and generally remaining the safest and most efficient operation by not altering generation levels below maximum output. On the other hand, some types of generation, such as hydro and natural gas, may be able to dispatch generation more often than their present operation – this represents flexibility in the current management scheme. We consider the effects of baseload generation in Section 3.5.

Further, we note that our assumption that renewables generation can be scaled from existing timeseries constrains the geographic diversity of the renewable generation in these grids at a fixed point. Though it would be possible to marginally improve the statistical multiplexing of these supply resources by considering wider geographical diversity, we feel that



(a) Summer (b) Winter

Figure 3.9: Five days of Ontario electricity blend scaled to 60% renewables annually in the (a) summer and (b) winter. Note that the grid does not maintain this level of renewables penetration throughout the year.

the existing sources of generation adequately represent the renewable generation resources available in each grid given current technology and economics.

3.4 Effects of Cost

Regulators, utilities, and independent energy generators consider a host of factors in deciding where, when, and what type of renewable energy generation to deploy. Beyond factors that affect energy output – such as site and transmission availability, weather patterns, and selection of generator technologies – there are a range of economic and policy factors. Weighing all of these inputs and making a rational choice is not straightforward. Tools like levelized cost-of-energy calculators [92] attempt to combine capital, operations and maintenance, performance, and fuel costs to provide a basis for comparison among different generation choices. However, these calculators often do not include costs related to financing, replacement, degradation, as well as possible benefits connected to subsidies and other policy drivers.

In this section, we study the sensitivity of Θ, the ratio of the 'costs' of deploying additional wind generation and additional solar generation. The cost here attempts to encapsulate all of the economic and policy considerations; we do not take a position on how to calculate this cost, but instead provide analysis to choose a mix that minimizes total solar and wind capacity given a particular relative cost, grid, and renewables penetration rate. Further, Θ

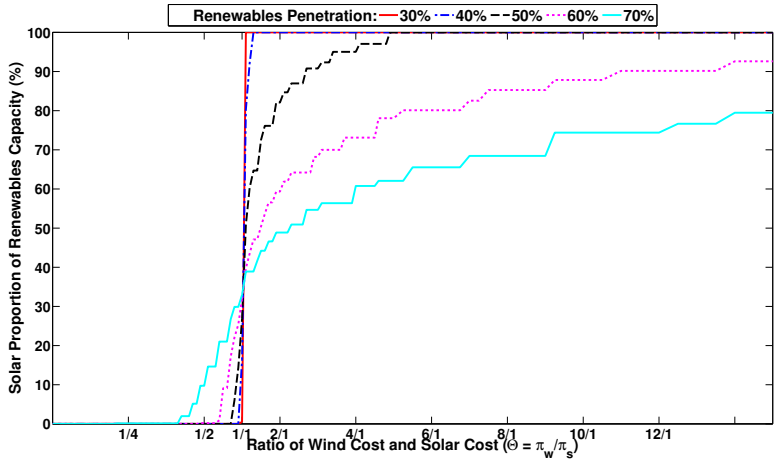


Figure 3.10: CAISO - California. Proportion of non-dispatchable capacity that is solar for different renewables penetrations and cost values. Wind is chosen in most scenarios, and required at high levels of renewables penetration ($p \geq 60\%$). This balance is reflected in the current non-dispatchable renewables generation mix in California.

may change over time, due to market factors and influence from policymakers. Additionally, our methodology applies to renewables technologies beyond those presented in this work, including variations of solar, wind, and tidal power. We note that for all the experiments in this section, there remains no baseload, non-displaceable generation ($\mathfrak{B} = 0$).

3.4.1 Effects of Cost on Renewables Mix

Figures 3.10, 3.11, and 3.12 show the solar proportion of the mix of non-dispatchable renewables capacity at various cost and renewables penetration levels for the California, Germany, and Ontario grids, respectively. The solar proportion here is the percentage of total non-dispatchable renewables comprised by solar generation by energy produced; thus, the remaining proportion is the wind proportion. These graphs result from proportionally scaling the non-dispatchable renewables sources found in each grid to create a range of penetration levels, and then varying the cost parameter Θ and minimizing the total cost at each point, as in Equation 3.1. A cost of $\Theta = 1/1$ indicates cost parity, and the left side of each graph ($\pi_s > 1$ and $\pi_w = 1$) represents higher-cost solar and the right side ($\pi_s = 1$ and $\pi_w > 1$) represents higher-cost wind. We note that the total capacity $a\mathfrak{C}_\mathfrak{s} + b\mathfrak{C}_\mathfrak{w}$ changes as a function of Θ and p – for more information, see Section 3.4.2.

In each case, at lower levels of renewables penetration, there is a value of $\Theta = \Theta^*$ across which there is complete substitution of one technology for another. This is because, at lower penetration levels, demand far outstrips renewable supply, and therefore the renewable supply need not match the pattern of consumption. Θ^* reflects the ratio of capacity factors of the technologies in each grid; in the case of Germany and Ontario, wind has a relatively

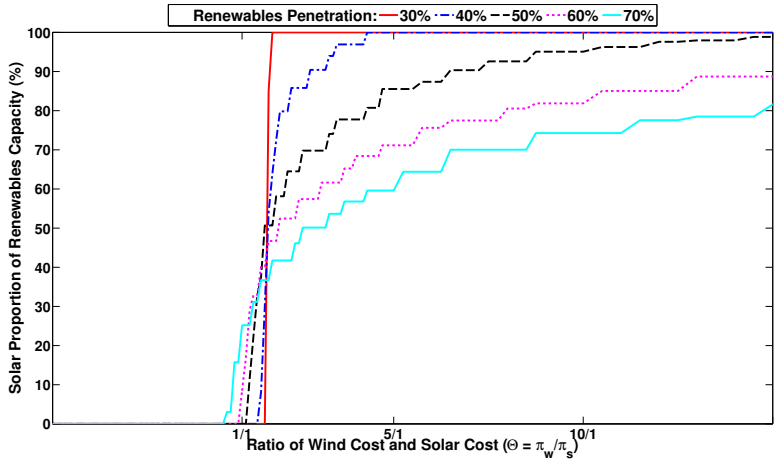


Figure 3.11: EEX - Germany. Proportion of non-dispatchable capacity that is solar for different renewables penetrations and cost values. Wind is required to meet renewables penetration levels $p \geq 40\%$. At cost parity ($\Theta = 1/1$), an all-wind blend is always chosen, differing substantially from the roughly 50%-50% mix that is the reality on the German grid.

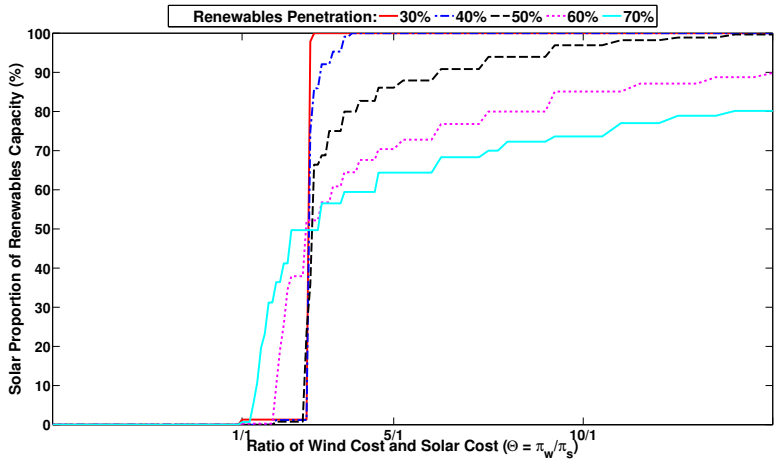


Figure 3.12: IESO - Ontario. Proportion of non-dispatchable capacity that is solar for different renewables penetrations and cost values. Wind is required to meet renewables penetration levels $p \geq 50\%$. At $p = 40\%$, the cost of wind must be very high ($\pi_w > 10$) to switch from a mixed blend to an all-solar blend.

higher capacity factor than solar, whereas in California, capacity factors are closer to parity.

For any penetration where $\pi_s \gg \pi_w$, the chosen mix is all wind; since wind generation, unlike solar, produces power at all hours, an all-wind mix can attain high penetrations. However, when $\pi_w \gg \pi_s$, a mix of both solar and wind is required to reach higher penetrations, as solar does not produce during the night hours but aligns better with daytime peaks in

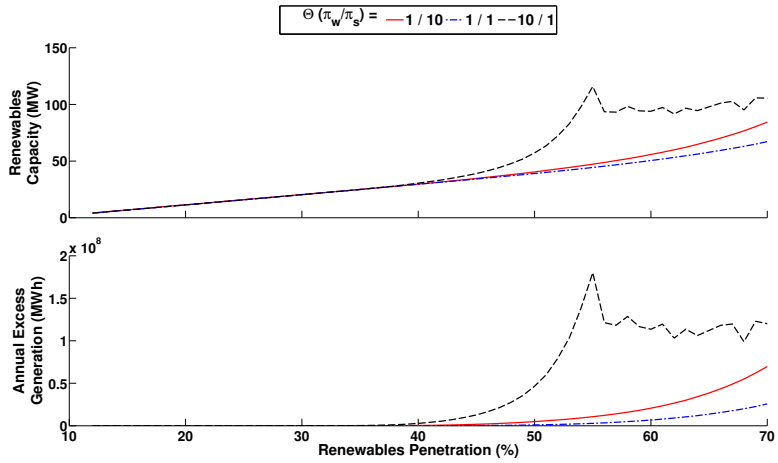


Figure 3.13: CAISO - California. Renewables capacity and excess versus renewables penetration at three cost points. The primarily solar blend ($\Theta = 10/1$) incorporates wind generation at $p > 55\%$.

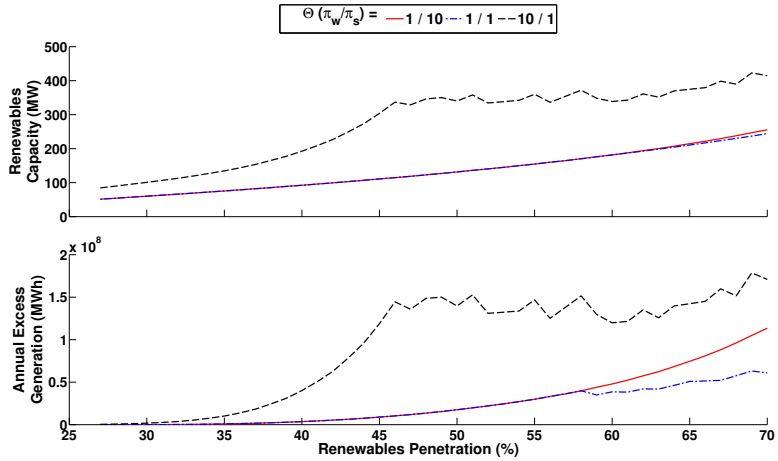


Figure 3.14: EEX - Germany. Renewables capacity and excess versus renewables penetration at three cost points. Since wind has a superior capacity factor, the cost parity scenario ($\Theta = 1/1$) chooses the same blend as the scenario where wind is favored ($\Theta = 1/10$).

demand. For each grid, there is a maximum penetration p_s for an all-solar mix. For example, p_s is between 50% and 60% for CAISO. Characterization of all-solar blends and p_s is further explored in Section 3.6.

Last, differences in curve shapes at a particular penetration level on different grids reflect the affinity of each type of generation for that grid's demand pattern; for example, if the curves are shallower, as in Germany and Ontario, more wind is required because the solar resource is not as potent. Also, each graph has a convergence cost Θ_C where, regardless of p,

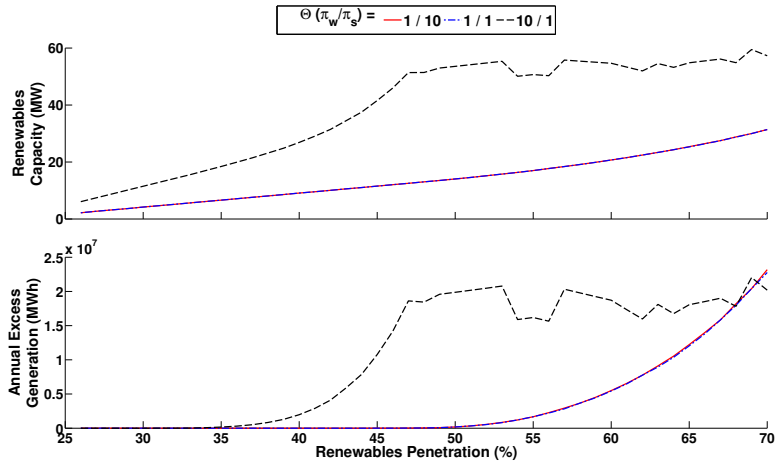


Figure 3.15: IESO - Ontario. Renewables capacity and excess versus renewables penetration at three cost points. Since wind has a superior capacity factor, the cost parity scenario ($\Theta = 1/1$) chooses the same blend as the scenario where wind is favored ($\Theta = 1/10$). For $\Theta = 10/1$ and $p \geq 40\%$, a mixed solar/wind blend is selected instead of an all-solar blend.

the ratio of capacities $\frac{a\mathfrak{C}_s}{b\mathfrak{C}_w}$ is roughly equal. At Θ_C, the technologies are in balance; further deployment should yield the same proportion of technologies as already exists, regardless of the level of renewables penetration.

Comparing these results to the current state of each grid, we can start to understand what current values of Θ might be. In the case of California, there is currently roughly eight times more wind capacity than solar capacity but the energy production of the two resources are approximately on par. This indicates that to this point, the cost of solar is greater than the cost of wind, or $\pi_s > \pi_w$. Though market forces such as the enormous reduction in photovoltaic panel cost will almost certainly have an effect on Θ going forward, this finding squares with the intuition that utility-scale solar has been slow to develop because costs remain high. In Germany, current capacity for solar and wind are nearly on par, even though wind generation has a significantly higher capacity factor in the German context. This indicates that the cost of solar is much less than the cost of wind in Germany, $\pi_s \ll \pi_w$. Given the policies in place that aim to build a strong German presence in the worldwide solar equipment market, whereby distributed generation is significantly compensated via a feed-in tariff, this finding also matches intuition; judging only on energy production and in the absence of subsidies, we would not expect the German grid to have half its non-dispatchable renewables capacity come from solar generation, especially at current levels of renewables penetration. In the Ontario grid, where strong solar subsidies are not in place but wind maintains a substantial benefit in capacity factor over solar, we see that the market has spoken by not yet deploying any utility-scale solar generation. This indicates that the cost of incremental wind generation π_w is *at maximum* slightly more expensive than the cost of solar generation π_s, but could also be less expensive.

3.4.2 Effects of Cost on Renewables Capacity and Excess Generation

The relative cost of different renewables has a significant effect not only on the renewables blend, but also on total renewables capacity and excess generation. Selecting less efficient technologies because of lower costs results in needing more total capacity. To represent this, Figures 3.13, 3.14, and 3.15 show renewables capacity $a\mathfrak{C}_\mathfrak{s} + b\mathfrak{C}_\mathfrak{w}$ and excess generation X for the California, Germany, and Ontario grids for three values of Θ: 1/10 (more expensive solar), 1/1 (cost parity), and 10/1 (more expensive wind).

In each case, cost parity selects the renewables mix to best represent the energy resources available, resulting in minimal total renewables capacity; from this perspective, any Θ besides parity requires either the same or more capacity. In the case of both Germany and Ontario, wind has a substantially higher capacity factor than solar, such that at cost parity, the capacity mix is all wind generation. Further, in the cases where $\Theta = 10/1$ (more expensive wind), solar is used until a penetration level where the benefit from incremental solar generation is smaller than a tenth of the benefit of adding incremental wind generation, so additional capacity from that point is a mix of solar and wind generation. Of these three grids, with cost parity, only the California grid benefits from having a balance of solar and wind generation. When considering excess generation X, it is important to not only look at the magnitude of generation, but also the capacity required to produce that generation. With this in mind, in each case, the cost-parity blend produces the largest X *per renewables capacity*. Further, more excess from the same capacity may be a benefit; excess could be an opportunity to charge storage, shift loads from other times, or encourage new energy-agile industries.

3.5 Effects of Baseload Generation

In Sections 3.3 and 3.4, we operate under the assumption that any non-renewable generation source can be displaced ($\mathfrak{B} = 0$). However, technology, policy, and economic considerations may dictate otherwise; for example, if regulators are attempting to limit carbon emissions instead of solely promoting renewables, it may be preferable for renewables to only displace generation from carbon-producing generation sources. In this section, we consider the ramifications to our methodology of having generation that is non-displaceable.

We begin by considering the California CAISO electricity grid and selecting a value of $\Theta = 1/1$, cost parity, to allow for comparison among scenarios. Additionally, we initially choose $\mathfrak{B}$, the non-displaceable generation, equal to zero. As renewables are introduced to the grid, they first displace nuclear generation, followed by imports, thermal (natural gas), and then hydroelectric generation. Figure 3.16 shows the evolution of the energy mix as renewables penetration increases all the way to 90%; values for each type of generation are given as a percentage of the reference demand $\mathfrak{D}$, and excess values that are beyond the dimensions of the chart are notated with an arrow and a number. Excess generation is

greater than 10% beyond demand for $p > 69\%$.

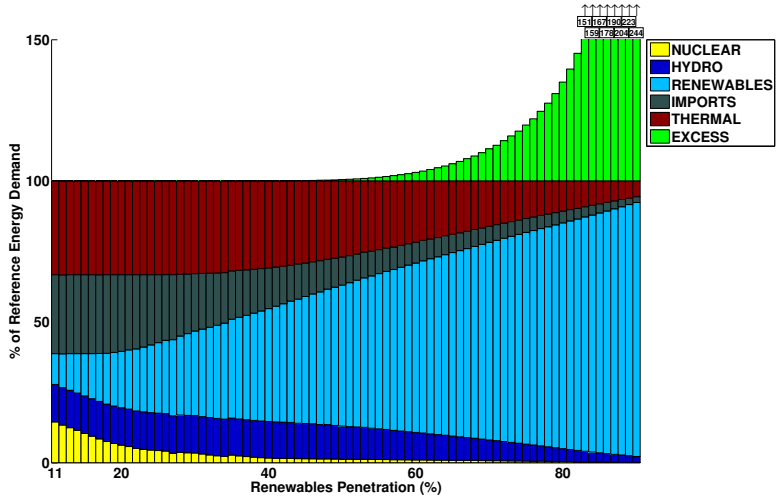


Figure 3.16: CAISO - California. Breakdown of supply versus reference energy demand at different renewables penetrations assuming that all generation is displaceable. Nuclear generation is displaced first, followed by imports, thermal, and then hydro.

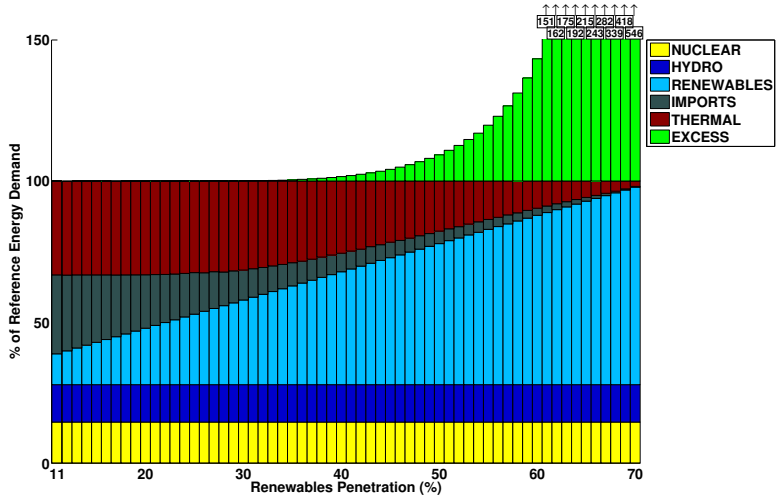


Figure 3.17: CAISO - California. Breakdown of supply versus reference energy demand at different renewables penetrations assuming Hydro and Nuclear generation are non-displaceable. Import generation is displaced first, followed by thermal.

Next, we observe how the mix changes when certain generation sources are non-displaceable. Whereas previously $\mathfrak{B}$ was zero, here $\mathfrak{B}$ contains all hydroelectric and nuclear generation, those classes that are not renewable by CAISO's definition but are carbon-free. Thus, as renewables are introduced to the grid, the first energy displaced is from imported generation, followed by that from thermal (natural gas) generation. Figure 3.17 shows the energy mix in

this scenario, though as $\mathfrak{B}$ now contains nearly 30% of generation, the chart has a maximum renewables penetration of 70%. Excess generation is greater than 10% beyond demand for $p > 50\%$, far less than the case where $\mathfrak{B} = 0$, and increases far more aggressively, with 546% of reference demand produced at $p = 70\%$.

As the total displaceable generation decreases, the ability of incremental renewables to displace remaining generation also decreases, resulting in diminishing returns. Another way to represent these diminishing returns is by looking at a chart representing the p-feasible mixes between solar and wind generation. In the earlier chart for California (Figure 3.1), the baseload generation $\mathfrak{B}$ was equal to zero. In this case, with $\mathfrak{B}$ equal to 28%, the resulting mixes are in Figure 3.18. It is evident that as p approaches 70%, the renewables capacity needed increases exponentially. Additionally, the excess generation $X(a, b)$ also increases exponentially; we examine the ramifications of the increase in excess energy in the following section.

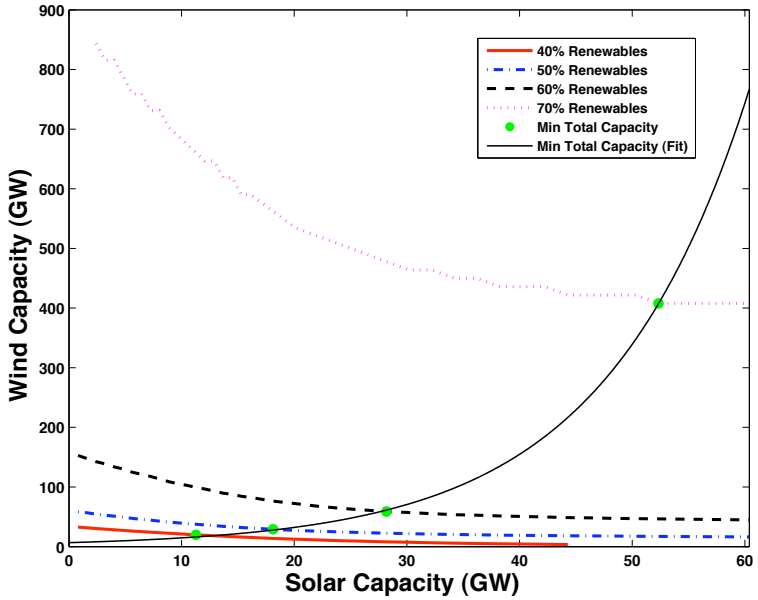


Figure 3.18: CAISO - California. Mix between wind and solar capacity at four renewables penetration levels assuming that Hydro and Nuclear generation are non-displaceable. Import generation is displaced first, followed by thermal. Note that minimum total capacity here assumes equal cost of incremental solar and wind generation.

3.6 Limits to Renewables Penetration

As grids incorporate more fluctuating renewable resources, a critical question is how much renewables capacity can be accommodated. Though the challenges of fluctuating renewables

are well known [89], we do not know of any study that studies limits to renewables penetration and their drivers.

Figure 3.19 presents excess generation X relative to demand $\mathfrak{D}$ in five scenarios from the CAISO grid, the EEX grid, and the IESO grid. Three of the scenarios, all with $\mathfrak{B} = 0$, are identical to those presented in Figures 3.10– 3.15. The additional scenarios include non-displaceable baseload generation $\mathfrak{B}$ preventing specific classes of generation like Nuclear and Hydroelectricity from being displaced. We do not present a scenario with nonzero $\mathfrak{B}$ for the EEX grid because German policy aims to eliminate nuclear generation. Additionally, we use a value of $\Theta = 1/1$, the mix of solar and wind generation at cost parity, to allow for comparison among results from different grids. Looking at differences among excess in these five scenarios, we identify two types of limits to renewables penetration: *hard limits*, insurmountable thresholds constrained either by the sum of $\mathfrak{R} + \mathfrak{U}$ or the total demand proportion during available renewable hours, and *soft limits*, arising from how much excess generation X the grid operator is willing to bear. Hard limits are represented by vertical asymptotes on the graph, while soft limits are thresholds on the graph relative to p and are a function of temporal patterns of renewables availability, the amount of non-displaceable generation ($\mathfrak{B}$), and the selected mix of renewables capacity ($a\mathfrak{C}_{\mathfrak{s}} + b\mathfrak{C}_{\mathfrak{w}}$). Soft limits are strongly influenced by the minimum value of p for which $X > 0$ and the rate at which X grows thereafter.

To better understand the ramifications of the renewables mix, Table 3.4 presents two additional scenarios for each of the five grid configurations: a solar-only grid and a wind-only grid. We can see a wide range of limits to renewables penetration, with soft limits spanning from near 30% all the way up past 90%, and hard limits extending all the way to 100%. A nonzero $\mathfrak{B}$ dictates that renewables have a lower hard limit, providing a narrower range of hours and smaller potential magnitude for renewables generation, resulting in even lower soft limits to penetration. We note that solar generation alone, because of its generation pattern, creates more excess at substantially lower renewables penetrations as compared to wind. All-wind blends tend to perform on par with cost-parity blends, as the generation mix is similar. In all cases, the cost-parity blend can accommodate the most renewables at the lowest generation, a reflection of the effect of Θ. The differences in the grids' blends with $\mathfrak{B} = 0$ shows that the wind generation on the IESO and EEX grids have better affinity to the load shapes on those grids, resulting in soft limits at higher renewables penetrations than the solar-wind CAISO blend.

3.7 Comparison to Related Work

Our work has much in common with other efforts in the literature. Here we consider three of these efforts: the SWITCH model for the western North America grid at UC Berkeley, work from Jacobson and Hart at Stanford, and the Renewable Energy Futures project from the National Renewable Energy Laboratory. These projects differ from our work in techniques employed, research goals, and grid scales examined; given this, we consider the relative

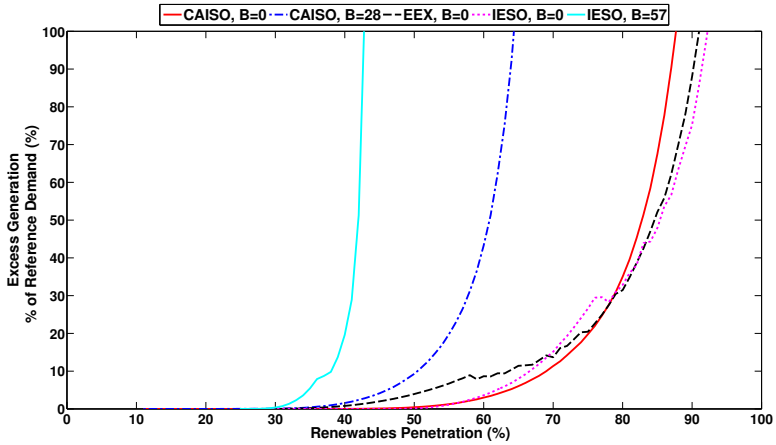


Figure 3.19: Excess generation on three electricity grids in scenarios with and without baseload generation. Each line is relative to that grid's total demand.

strengths and weaknesses of each project as compared to our work.

3.7.1 SWITCH

SWITCH [54] is a planning tool for generation, transmission, and storage infrastructure in western North America. With current infrastructure given as an input, SWITCH determines the additional infrastructure needed over a planning horizon, typically until 2030 or 2050, given a set of economic, technology, and pollcy constraints. SWITCH ensures that the given investment plan provides sufficient electrical generation over all the hours of its simulation, typically 144 hours distributed throughout the year per each investment period, which varies between 4 and 10 years in different studies. To perform the optimization, SWITCH uses a mixed-integer linear program that encodes the constraints and objective function to find the lowest-cost infrastructure plan. Once this plan is found, SWITCH runs again to add additional capacity to ensure suitable reserve margins, and then runs once more to calculate total installation, maintenance, and operation costs.

SWITCH has a number of strengths that go beyond the capabilities of our model. It considers a range of different intermittent renewables generation technologies; when choosing additional generation infrastructure, SWITCH considers a number of different generation facilities, including wind, solar photovoltaic, solar thermal, and combined-cycle natural gas. The model couples this new generation with existing nuclear, coal, and hydroelectric facilities and also considers reserve margins. Further, SWITCH optimizes decisions to meet electricity demand in regions rather than in aggregate. In connecting these regions, SWITCH uses a simplified model of transmission lines that is able to decide where additional infrastructure would be most useful.

The SWITCH model fares worse than our techniques in two areas. The sampling methodology for choosing hours against which to optimize planning decisions may result in inade-

quate resources for peak events. These challenging events may have a duration that exceeds the maximum six consecutive hour sample used by SWITCH, and further, these events are often consecutive themselves. Thus, the severity of peak events may be underestimated by the sampling methodology of SWITCH, resulting in higher estimates of available generation, transmission and storage resources. Our methodology addresses this by testing over the consecutive hours of an entire year, capturing the full time-series of operation. Also, SWITCH gives minimal consideration to demand-side flexibility [53]. Load flexibility is encoded by incorporating price-sensitive demand on an annual basis, meaning that higher electricity costs result in a fixed portion of demand that responds by reducing reserve margins for the duration of the year. On the other hand, our approach looks at the potential of elastic demand on an hourly basis.

3.7.2 Jacobson/Hart

Jacobson and Hart [60] produce a model of the California electricity grid by creating generation functions for each intermittent technology – wind, concentrating solar power (CSP), and rooftop photovoltaic (PV). The generation functions are constructed from two years of meteorological data and resource availability models. They also incorporate data from other facilities, including hydroelectric, nuclear, geothermal, and natural gas. For intermittent sources, generation functions are used to perform Monte Carlo simulations coupled with generation from other sources to produce least-cost or least-emission dispatch plans. Each generation type also has technology-specific curtailment, ramping, and dispatch rules.

This work tests each technology in isolation, evaluating its effectiveness at reducing the need for natural gas generation and characterizing its "curtailment point" – the maximum capacity at which none of its generation is curtailed. They also consider the maximum theoretical penetration of each technology, akin to our classification of limits of renewables penetration. Then, by using those results, the authors find the balance of solar and wind technologies that is most successful at carbon abatement. Also similar to our work, the authors note that in a scenario with substantial renewables, the critical periods are different than the present-day grid, often occurring at night or during autumn. The authors' approach incorporating ramp rates and dispatch limitations of each technology are a particular strength relative to other work.

However, similar to the SWITCH study, the strategy used for modeling the entire grid may result in underestimating the severity of peak events. By doing Monte Carlo simulations, a statistical sample of the distribution of the generation functions is considered, though the grid must be able to respond to outlier events, not only common or even infrequent events. Also, the authors mention storage and flexible demand via demand response capability but do not analyze their potential to displace other types of generation.

3.7.3 NREL RE Futures

The Renewable Energy Futures study by the National Renewables Energy Laboratory [93] contains the most complete high-renewables United States electricity grid planning model available, to our knowledge. It combines the Regional Energy Deployment System (ReEDS) linear programming model [109], which has a similar construction to SWITCH but considers generation and transmission for the entire contiguous United States, with the GridView model from ABB [1], which provides unit commitment on an hourly basis to evaluate the investment plans laid out by ReEDS. Together, these two systems evaluate a range of renewables scenarios for 2050 incorporating different policy, technology, and economic assumptions.

A primary strength of the RE Futures study is the wide range of generation, storage, and flexible demand technologies considered. The renewables generation types considered are onshore and offshore wind, CSP with and without thermal storage, utility-scale and distributed PV, dedicated and co-fired with coal biomass, geothermal, and hydroelectric. For storage, pumped hydroelectric, compressed air, and batteries are considered. On the demand side, interruptible load, building thermal energy, and utility-controlled plug-in electric vehicle charging are evaluated. This wide array of carbon mitigation technologies yields an enormous state space and a far more flexible model than any other in the literature. In particular, the treatment of flexible loads and evaluation of distributed solar PV (using the NREL Solar Deployment Systems (SolarDS) [43]) in the report are unmatched in other holistic electricity grid models.

One area in which RE Futures could have been improved is its sampling of grid demand. Each year in the study is represented by 17 time slices, composed of four slices per season plus one additional slice to represent the summer demand peak. This sampling methodology was used to limit the computational complexity of the model, but, just as in the other models, may result in insufficient estimation of the severity of peak events. By using a continuous dataset, our work aims to avoid this shortcoming. Our work expands on RE Futures by considering varying levels of implementations of demand-side management technologies, rather than a small number of discrete scenarios.

Renewables Penetrations						
California - CAISO						
	All Displaceable $\mathfrak{B} = 0$			**Hydro & Nuclear Non-Displaceable $\mathfrak{B} = 28\%$**		
Renewables Mix	**Solar Only**	**Wind Only**	**S/W Mix, Θ=1/1**	**Solar Only**	**Wind Only**	**S/W Mix, Θ=1/1**
10% Excess	46.5%	60.7%	69.0%	35.6%	44.2%	50.4%
25% Excess	50.0%	68.2%	77.1%	38.3%	50.1%	56.6%
50% Excess	52.5%	74.1%	82.8%	40.1%	54.6%	60.8%
100% Excess	55.1%	79.7%	87.7%	42.1%	58.8%	64.3%
Hard Limit	63.5%	100.0%	100.0%	47.1%	72.2%	72.2%
Germany - EEX						
	All Displaceable $\mathfrak{B} = 0$					
Renewables Mix	**Solar Only**	**Wind Only**	**S/W Mix, Θ=1/1**			
10% Excess	40.4%	59.2%	63.6%			
25% Excess	44.7%	69.8%	77.1%			
50% Excess	48.1%	77.8%	84.6%			
100% Excess	51.5%	85.6%	91.0%			
Hard Limit	66.6%	100.0%	100.0%			
Ontario - IESO						
	All Displaceable $\mathfrak{B} = 0$			**Nuclear Non-Displaceable $\mathfrak{B} = 57\%$**		
Renewables Mix	**Solar Only**	**Wind Only**	**S/W Mix, Θ=1/1**	**Solar Only**	**Wind Only**	**S/W Mix, Θ=1/1**
10% Excess	46.2%	66.6%	66.6%	31.5%	40.1%	38.0%
25% Excess	49.9%	74.2%	74.4%	32.1%	41.1%	40.6%
50% Excess	52.9%	81.2%	85.4%	32.4%	41.9%	41.9%
100% Excess	55.6%	88.1%	92.2%	32.9%	42.4%	42.8%
Hard Limit	67.7%	100.0%	100.0%	34.9%	43.4%	43.4%

Table 3.4: Hard and soft limits to renewables penetration in scenarios on the California, Germany, and Ontario grids. The level of excess generation allowed by the grid operator represents soft limits on renewables penetration, whereas fundamental limits on resource availability and energy demand constitute hard limits on renewables penetration.

Chapter 4

Increasing Limits to Renewables Penetration

In the previous chapter, we established a methodology for varying the penetration of renewables on electricity grids and, using that methodology, arrived at a formulation for limits to renewables penetration on electricity grids. In this chapter, we explore some critical techniques relative to the temporal dynamics of our year in a grid with deep renewables for increasing limits to renewables penetration: the coordinated management of the entire portfolio, the utilization of storage resources, the potential to modulate (or dispatch) demand, and grid-driven demand reduction. We evaluate the efficacy of these techniques for creating a better match between supply and demand by considering their impact on reductions in renewables as well as fossil fuel generation, encouraging an increase in limits to renewables. Ultimately, all of these aspects need to come together in a manner that addresses the additional level of fidelity associated with transmission constraints, plant dynamics, demand adjustment mechanisms, and markets.

Here, we perform all evaluations using the California CAISO grid, and we make the assumption that baseload non-displaceable generation includes nuclear and hydroelectric facilities (*i.e.*, $\mathfrak{B} = 28\%$). We make this assumption because of the practical considerations that these types of generation represent enormous investments unlikely to be stranded, are carbon-free and therefore less likely to be replaced on policy grounds, and are fairly mature generation technologies with limited potential for innovation.

A sustainable grid presents a family of challenges associated with coordinating heterogeneous, distributed processes to better manage increasingly critical dispatchable resources. For instance, consider the duration curves in Figure 4.1 – the present-day thermal curve, representing current use of natural gas generation, indicates the traditional need for critical peak demand response. A large fraction of dispatchable capacity is utilized only a tiny fraction of the time, whereas other generation is used nearly continuously. By curtailing or shifting load away from the critical times where expensive yet underused generation is employed, capital investment in generation can be greatly reduced. However, the impact of this volatile minority of the portfolio is amortized over the large body of supplies that are

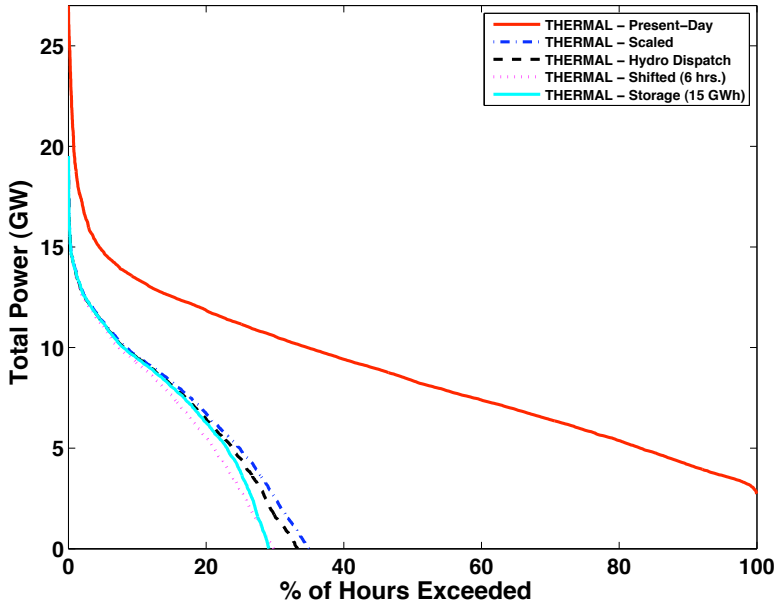


Figure 4.1: Comparison of duration curves for thermal generation before and after scaling, as well as after supply or demand modifications.

essentially providing baseload and seasonal adjustment.

In a sustainable grid, this dispatchable resource takes on a new and critical role of providing firming to renewables, which, recall from Figure 2.4, are more variable and less predictable than grid demand. The nature of this can be seen in the thermal curve for the 60% renewable "scaled" blend in Figure 4.1. Most of the time, these resources are unused, but the reduction in peak, from 27 GW to 19 GW (approximately 30% of capacity), is less than the reduction in overall usage (from 33.3% of total energy to 9.7% of total energy, a reduction of 71%). Further, the upper 25% of resources is used a tiny fraction of the time. More sophisticated coordination algorithms would use storage, load shifting, and curtailment to eliminate this sharp peak, even if mean usage of these resources increases marginally. The control regime of these dispatchable resources is likely to differ markedly from today, because it is not enough to only modulate reserve to preserve power quality, instead ramping some resources all the way off and back on becomes essential. However, the frequency of these transitions needs to be minimized for each individual plant.

Yet another constraint on sustainable grids is the agility of dispatchable generation; a distribution of the hourly changes experienced by thermal generation is in Figure 4.2. Though the most common "change" in the scaled grid is zero, the changes have a longer-tailed distribution than those in the present-day grid. We cannot ascertain whether increased changes are feasible, as CAISO does not release data on allowable ramp rates of the generation fleet. Nonetheless, we emphasize that the dispatchability and agility that fossil fuels provide are critical to operating a grid composed primarily of renewables. This presents a new role for

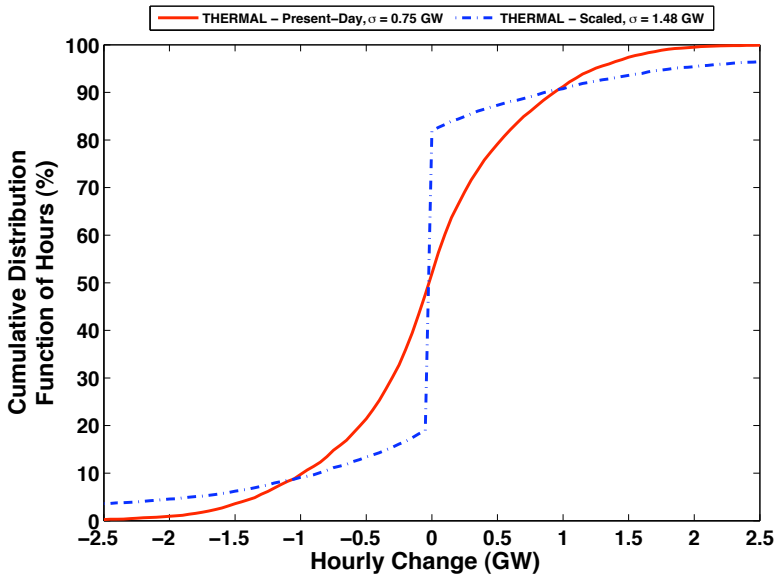


Figure 4.2: California - CAISO. Comparison of distributions of hourly change for thermal generation before and after scaling. Increased variability in renewables results in less often but higher magnitude dispatch of fossil fuel resources.

fossil fuels: rather than providing baseload power and seasonal adjustments, fossil fuels are a precious storage resource that can be used only once, providing much-needed firming to cope with deep uncertainty in both supply and demand.

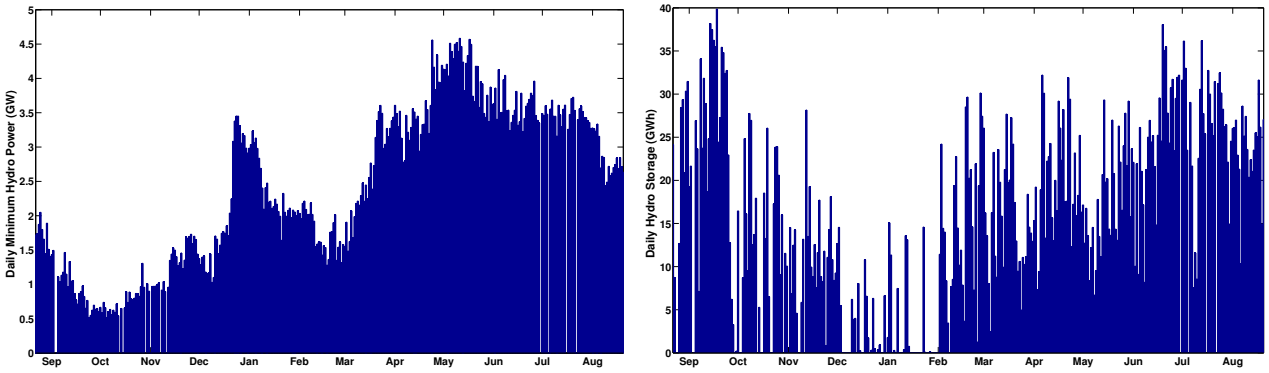
In this chapter, we modify the operation of supply, storage, and demand resources for improving the match between supply and demand. Though we evaluate our changes primarily on the reduction in the magnitude of natural gas generation needed to operate such grids, we instead could have considered how these modifications reduced renewables capacity. Further, with generation data that is more granular (*i.e.*, by facility), we might consider operation of specific facilities in our evaluation, but in the absence of such data, we consider generation aggregated by type.

4.1 Supply-Side Management

As the portion of generation capacity that is non-dispatchable becomes larger, the remaining portions that are dispatchable become more valuable. To assess this value, we examine the potential for improved coordination of these disparate generation resources, including those traditionally used for baseload power. As an example of the value of systems co-optimization on the California electricity grid, we examine the large hydroelectric resource, comprised of facilities with capacity over 30 MW. Looking at the daily patterns in hydroelectric production in Figure 3.7, it is clear that operators are dispatching these resources daily

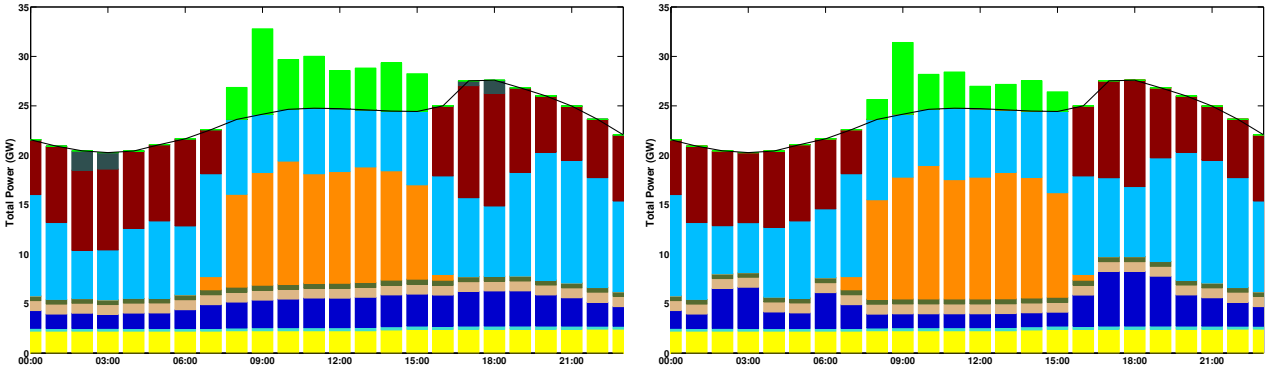
to perform load-following, matching supply to overall demand. To quantify the magnitude of this dispatch, we identify the minimum for each day, defining it as a daily "floor" of hydroelectric power production. We also identify the annual "ceiling" – the most power delivered above the daily floor by the hydroelectric category over the entire year, which is 4111 MW. This ceiling limits the algorithm from ever calling for more than the maximum achieved flexible generation over the year. The sum of the generation above this floor over all of the hours of the day constitutes the capacity of this dispatchable resource, reflecting the volume of water than can be opportunistically released. Figure 4.3(a) shows the daily minimum hydro power, representing the floor values for each day. Figure 4.3(b) shows the daily availability of the flexible hydroelectric resource for a year. As a check on this methodology, the magnitudes of both the floor and the flexible dispatch vary by season – they peak in the summer and reduce thereafter, coinciding with the availability of snowmelt from the Sierra Nevada.

If the objective of this dispatch were altered from load-following to firming shortfalls of supply relative to demand, this generation could provide substantial zero-emissions load balancing capacity. Using a simple omniscient algorithm that reassigns dispatch from this resource to first remove imported and then thermal generation, we attempt to quantify the utility of this resource. Though the total volume of water that is dispatched for generation is large, only a subset of this water can be used as dispatch able generation, as the hydroelectric generation is being used to meet existing demand. Thus, the only hydroelectric generation that can be shifted occurs when there is excess generation that can make up for the reduction in hydroelectric generation used for shifting. Using this condition, we construct a model of a grid that uses the flexible hydroelectric management resource described. We can see operation of this model in Figure 4.4, showing a single day before and after modifying management of the hydroelectric resource. By employing this change, we minimize use of this valuable load balancing resource when non-dispatchable renewables are providing excess generation, and increase use of the load balancing capability when natural gas plants would otherwise be dispatched. Looking at the capability of this resource over an entire year in Figure 4.5, we can clearly see how the availability of this flexible resource varies throughout the year, providing little to no capacity in the winter months and over 1 GW of capacity in the summer and autumn. Over the year, this method reduces the total thermal and imported energy by 6%, though does not reduce the peak thermal energy because excess is not available at that time. To further understand the effect of this approach on the use of natural gas on this grid, Figure 4.1 compares this method to the 60% scaled blend with no modification. This shows that this method achieves a modest reduction in the fraction of hours where natural gas is needed but has limited potential to target the hours with the highest magnitudes. It may be possible to improve the ability of these dispatchable generation sources to displace peak natural gas generation by *increasing* natural gas generation at certain hours. For example, additional natural gas generation at the non-peak hours of the day could be used to displace some hydroelectric generation, allowing for that generation to be applied during the peak natural gas generation hours. The incremental addition in some hours allowing for reduction in other, more critical hours helps to achieve



(a) Daily Minimum Hydro Power Delivery (b) Daily Total Hydro Capacity

Figure 4.3: CAISO - California. Daily minimum hydroelectric power and daily total energy available for portfolio-aware hydroelectric management. This represents the annual pattern of hydroelectric availability and the total energy flexed above the minimum delivered power each day.



(a) Grid Scaled to 60% Renewables (b) Scaled Grid with Hydroelectric Management

Figure 4.4: CAISO - California. An example day before and after employing portfolio-aware hydroelectric management.

the goal of flattening the duration curve, despite using more natural gas generation overall.

The shift from using this dispatchable hydroelectric resource to address peaks in the demand curve to using the same resources to reduce fossil fuel needs can apply to other generation types as well. For example, geothermal, biomass, and biogas facilities all may have flexibility in generation schedules. Like natural gas facilities, these facilities have their own capacity, ramp rate, and capacity factor constraints to consider.

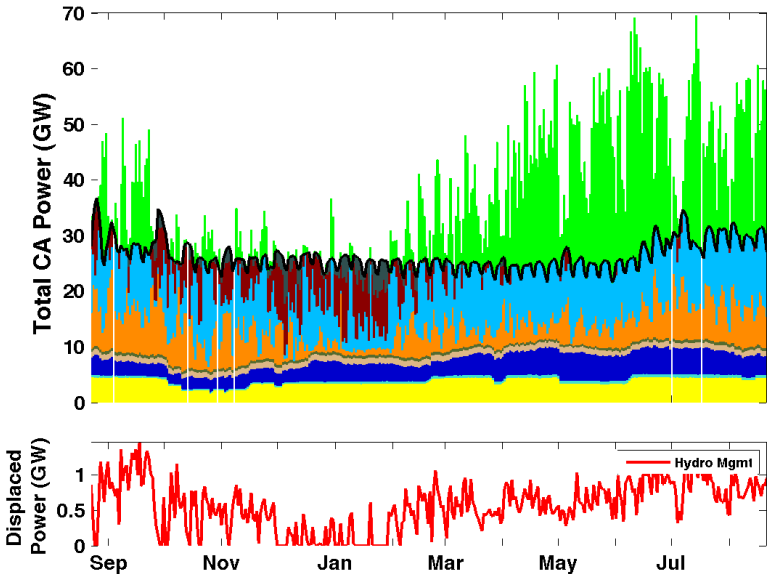
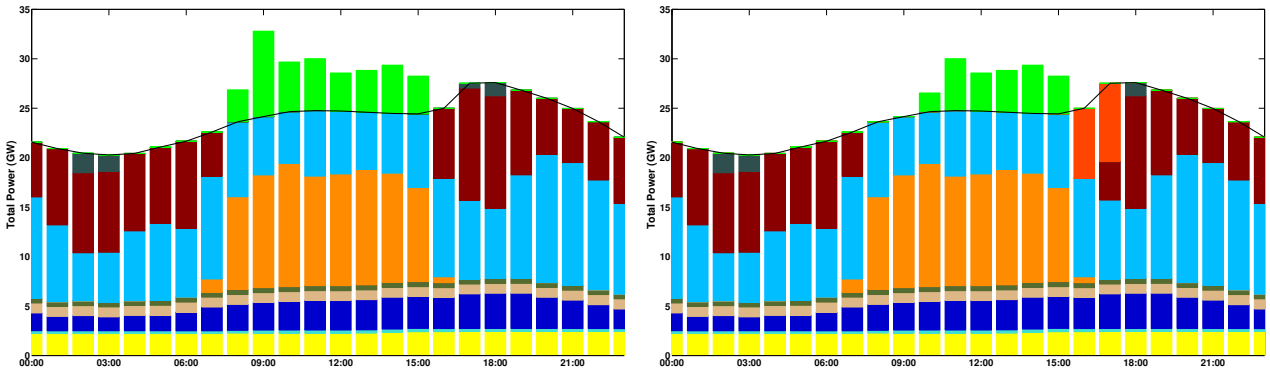


Figure 4.5: CAISO - California. Timeseries of 60% scaled grid operation and maximum usage of portfolio-aware hydroelectric management resource.

4.2 Grid-Scale Energy Storage

Another critical resource for improving the supply-demand match is grid-scale energy storage. Currently, 128 *GW* of storage capacity is already deployed throughout the world (about 6 *GW* in California), but over 99.9% of it is pumped hydroelectric storage [49, 104], where water is pumped from lower- to higher-altitude reservoirs when electricity is cheap and later released through generation turbines when it is expensive. Though other, more dynamic storage media such as compressed air caverns, flywheels, and novel battery chemistries are in development [105], most require additional maturation to be viable at scale.

Putting aside these technological challenges, opportunities for improved management of grid-scale energy storage lie in modeling and managing storage to reduce variability and maximize utilization in the unpredictable and dynamic landscape of a renewables-heavy grid. Though storage management is conceptually similar to that of load management, there are key differences. Energy storage allows shaping of supply rather than demand, and is substitutable for any type of generation – unlike demand, where loads are unique, any unit of supplied power is equally useful. However, to provide this capability, storage bears a roundtrip efficiency factor, known to be 75% for pumped hydro [88]. Nonetheless, the ability to shift supply becomes increasingly valuable in a blend with deep renewables. To demonstrate this, we model a rudimentary storage resource that charges whenever excess is available and discharges without loss in place of imported and thermal generation whenever capacity is available. Figure 4.6 shows a single day before and after employing this storage resource on a 60% scaled CAISO grid. We see that excess is used as it is available, but



(a) Grid Scaled to 60% Renewables

(b) Scaled Grid with 15 GWh Storage

Figure 4.6: CAISO - California. An example day before and after employing grid-scale energy storage.

additional excess beyond the capacity of our storage resource is not used; in this case, a total of 15 GWh does not get used. Once natural gas and imported generation is needed after the period of excess, the storage resource empties, with no limitation in power delivery. Though this idealized model makes significant assumptions about the behavior of this storage resource, it provides an estimation of the value of this resource. Figure 4.7 provides a timeseries of the use of this storage resource over the year. In this case, the resource is often used in the fall and winter to displace natural gas, but sees little use in the summer when natural gas is seldom needed. Over the year, this method reduces the total thermal and imported energy by nearly 8%, though does not reduce the peak thermal energy because excess is not available to charge the storage resource prior to that time. Looking at the effect of this 15 GWh storage resource on the natural gas duration curve in Figure 4.1, it shows even further improvement on the fraction of hours where natural gas generation is needed, though many of these hours may be the same as those addressed by hydroelectric management improvements. This presents an opportunity for algorithm development, centering on choosing which of a set of resources with different strengths and weaknesses to employ with current and predicted conditions; we leave this to future work.

A grid-scale storage resource as described in this section provides operators with a valuable agile tool for improving the supply and demand match, but there is significant room for improvement, including more realistic power delivery limitations, use of a roundtrip efficiency factor, and more intelligent, non-greedy algorithm to decide when and how much to charge and discharge. One important consideration is that this storage only charges when excess energy is present; in fact, it may make sense to charge this storage resource using increased thermal generation at times. A more advanced storage controller could better target the peak and flatten the demand curve as well as improve utilization of this valuable and potentially expensive storage resource.

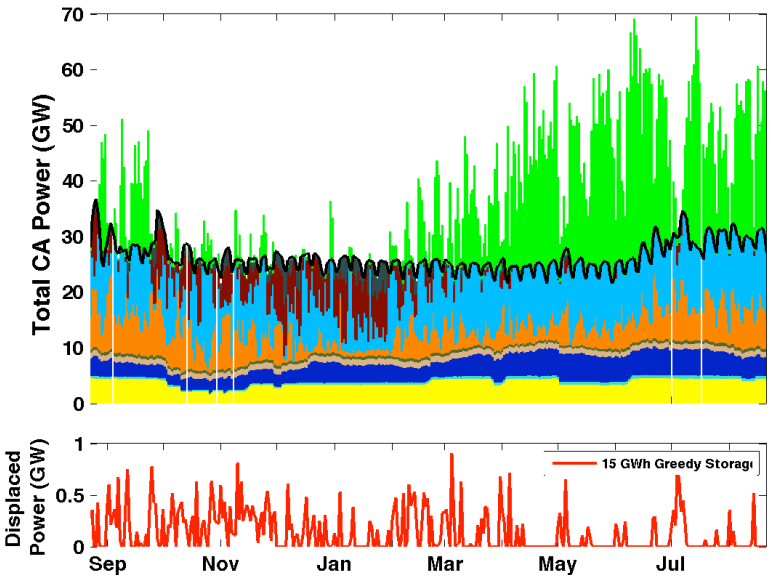


Figure 4.7: CAISO - California. Timeseries of 60% scaled grid operation and maximum usage of 15 GWh grid-scale storage resource.

4.3 Demand-Side Management

Aside from the modest gains possible via modifying the operation of supply-side generation resources and the enormous investment needed for largely unproven storage infrastructure, there is substantial potential to improve the match between supply and demand by utilizing existing demand in different ways. In this section, we examine supply-following loads, where certain loads can be advanced or deferred to operate when supply is available, and energy curtailment, where certain loads can be avoided when supply is in critical deficit.

4.3.1 Supply-Following Loads

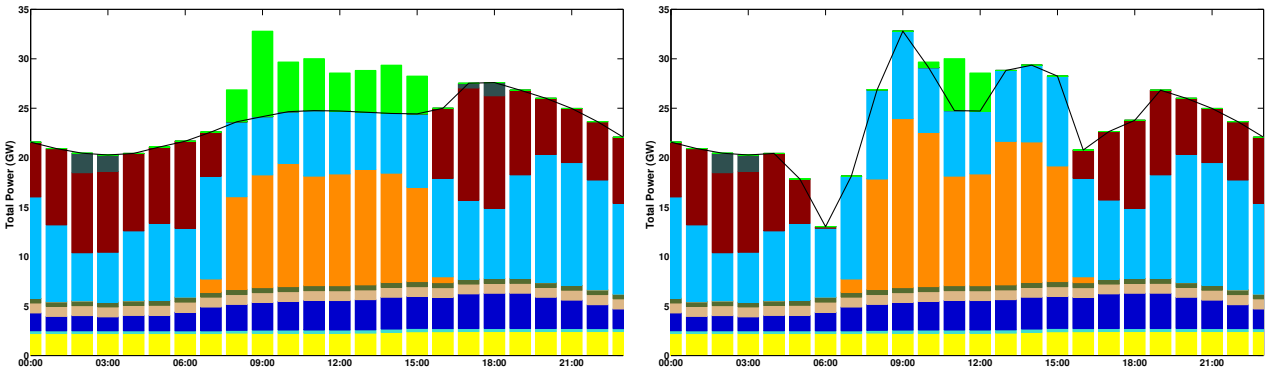
Existing grid architectures are predicated on the notion that operators can control generation supplies in order to match demand (load-following), but that while demand is predictable, the ability to shape it (its *elasticity*) is limited. Deep penetration of fluctuating renewables presents the dual challenge that the fraction of supply resources that are controllable is reduced and those resources are needed to firm renewables in addition to following demand. With increased communication and intelligence, it is particularly attractive to shift part of the burden of maintaining the match onto the demand side, i.e., to create a class of *supply-following loads*. A body of research work has explored such mechanisms for a wide variety of loads, as discussed in Section 4.5. Here we seek to quantify the potential gain of such approaches and to characterize the trade-offs in duration, magnitude, and agility that can guide their development.

Conceptually, a supply-following load is a mechanism that can shift demand away from periods of energy deficit to periods of energy surplus. For example, in Figure 3.7(b), shifting would allow loads adjacent to periods of excess to advance or delay operation in order to make use of renewable generation, thereby reducing imported and thermal generation. Periods with more excess have more potential for shifting, and increasing the time horizon allows more loads to shift away from fossil fuel-based generation.

To characterize the potential gain of such shifting as a function of the length of the shifting time horizon, we run a fairly aggressive shifting algorithm over the scaled trace. A key observation is that the opportunities for shifting occur near the crossover points of excess and deficit. First, a list of all possible shifting opportunities is constructed, containing "candidate" hours where fossil generation is used that are within the shifting time horizon of a "target" hour with excess generation. This list may contain multiple target options for a candidate, as well as multiple candidates vying for a target. Beginning with candidates that have to shift the longest number of hours, target hours are matched to candidates by identifying candidates that only have one potential target hour. The lesser of the candidate and target amount is shifted to displace the excess; if the excess is eliminated, the target is removed. Ties are broken by favoring delaying loads over advancing loads. This process is iterated until there are no more (target, candidate) pairs. We show a single day before and after shifting loads by up to three hours in Figure 4.8. We also show the yearly implications of up to 6 hours of shifting in Figure 4.9. Over the year, this method reduces the total thermal and imported energy by over 15%, though does not reduce the peak thermal energy because there is no excess on the day of maximum thermal generation and thus no target hours. Similar to the storage resource, the controller only employs load shifting to address periods of deficit when natural gas generation is otherwise used. However, the potential to both advance and defer loads improves the performance of this resource versus storage. Figure 4.1 compares the supply management, storage, and load shifting strategies; the load shifting strategy shows the most significant improvement in supply-demand matching throughout the duration curve of natural gas generation. Though the number of hours where natural gas generation is needed is slightly more than the energy storage scenario, the reduction achieved better targets the more challenging hours where natural gas generation is larger.

This method is optimistic because (1) decisions to shift loads are made *post hoc* given full knowledge of future loads and (2) demands and loads are advanced or delayed with no change in energy consumed. It provides a reasonable upper bound on the potential for load shifting capabilities. It also serves as a target for more practical real-time algorithms that utilize prediction of demand and supply conditions. The second assumption reflects the difficulty in making realistic claims about the shiftability of a diverse class of loads in a top-down analysis. A more detailed shifting model, *e.g.*, for shifting of thermostatic loads by precooling, might incorporate a penalty for shifting resulting from the physical processes involved (*e.g.*, heat transfer).

To assess the overall potential for load shifting to displace natural gas generation, we conducted experiments to vary the hours of shifting and renewables penetration limits; blends are computed using the method described in Section 3.1. Figure 3.4 reveals the fundamental



(a) Grid Scaled to 60% Renewables (b) Scaled Grid with 3 Hrs. Load Shifting

Figure 4.8: CAISO - California. An example day before and after employing load shifting.

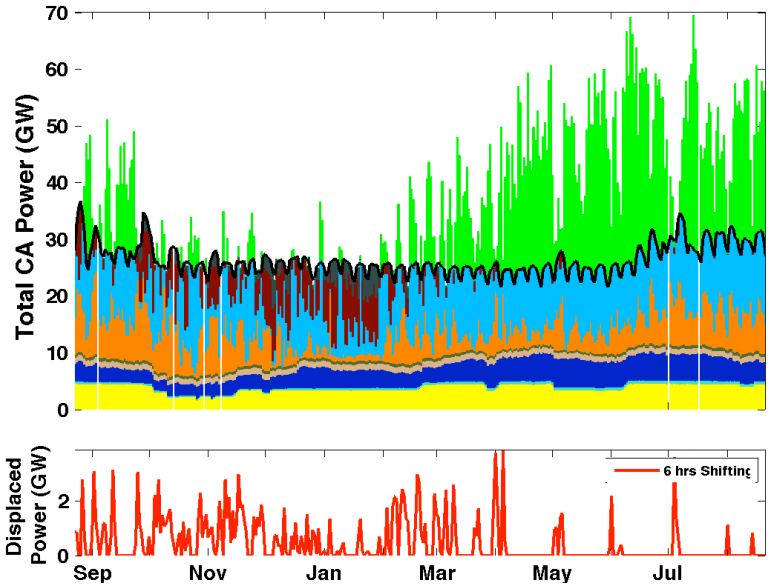


Figure 4.9: CAISO - California. Timeseries of 60% scaled grid operation and maximum usage of ±6 hours of load shifting resource.

limits of shifting: it has limited impact in summer when renewable supply is abundant, and in winter when it is scarce. It is most effective in the shoulder seasons where there are intermittent periods of excess alternated with periods of fossil fuel generation. Figure 4.10 shows the gain throughout the full year over a range of renewables penetration and shifting durations. The value of shifting increases strongly with the penetration – the number of target hours where excess energy is present increases faster than the decrease in the number of hours where fossil fuels are used for generation. Since there is more excess to utilize and less fossil to displace, it will be much more important in future more sustainable grids than

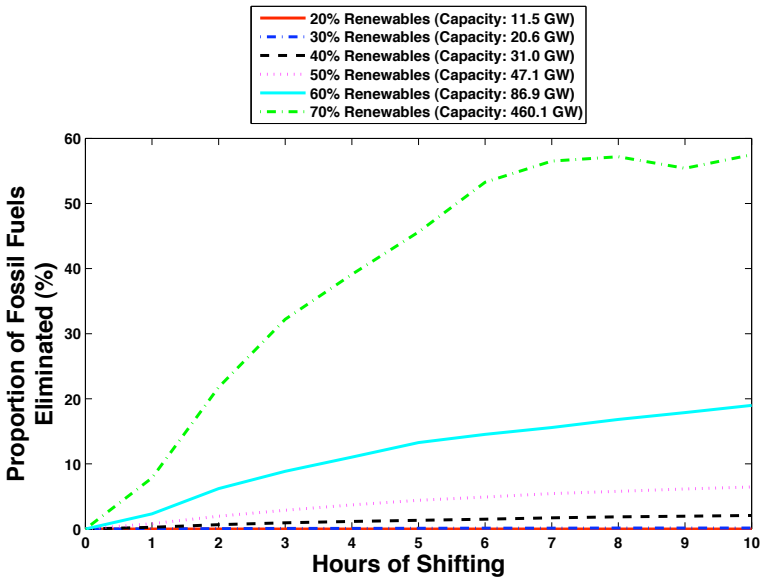


Figure 4.10: California - CAISO. The effect of load shifting on the fossil fuels in the generation blend. Note that as renewables penetration increases, the absolute percentage of fossil fuels in the blend decreases.

today. However, the overall impact of shifting is fundamentally limited by seasonal effects.

The most exciting aspects of supply-following loads lie in going beyond the simplistic algorithms modeled here. Rather than merely minimize total fossil generation, it is essential to utilize shifting to flatten the duration curve for those supplies. By targeting peak hours of fossil generation, we arrive at a novel variant of peak shaving. In addition, the agility associated with shifting can be used to take advantage of those opportunities of excess.

4.3.2 Energy Curtailment

In a renewables-heavy grid, the critical periods change from summer peaks in energy demand to winter lulls in renewable energy generation. This elicits a concurrent change in strategies for load curtailment, situations in which grid operators target the hours that need the most expensive generation by selectively interrupting non-essential loads. In our analysis, these critical hours appear in the leftmost points of the duration curves in Figure 4.11 – this is when high-cost, very low-utilization generation resources are operated. When this demand cannot be shifted elsewhere on the duration curve using other mechanisms, load curtailment can avoid these costly peaks. For example, in the 60% scaled CAISO grid, the peak 1% of hours over the year for thermal generation each consume at least 14 GW and up to 19.5 GW – over 28% of the capacity. Though this long tail is well understood for electricity demand, the difficulty in predicting and responding to peak events in a more dynamic electricity system is a critical challenge for achieving sustainable grid operation. Further, since the peak in fossil

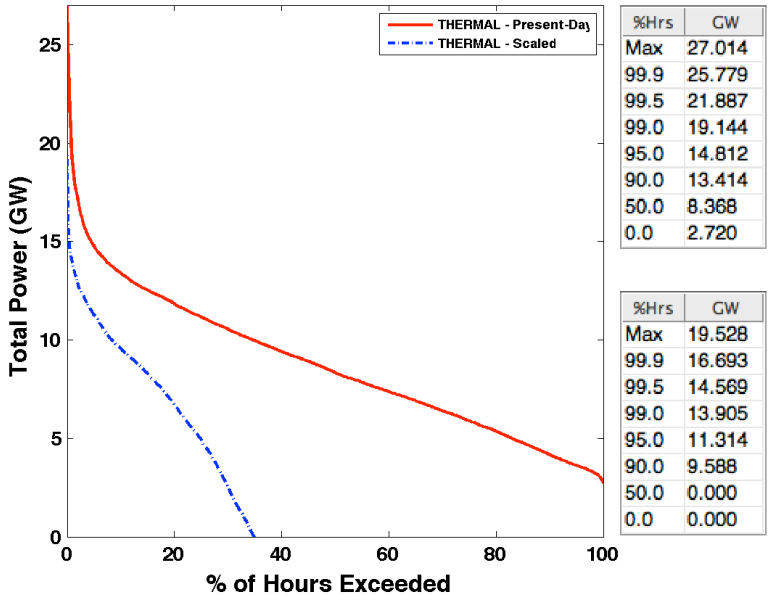


Figure 4.11: CAISO - California. Duration curves for themal generation in the unscaled and 60% scaled grids. A table provides selected points along each line.

fuel-based generation is no longer coincident with the peak in demand in the grid scaled to 60% renewables, there are more options for curtailment targets; unlike the status quo, curtailment can be most effective even when the grid is not at its most constrained overall.

4.4 Energy Efficiency

Energy efficiency is improved when less energy is consumed to provide the same service. In the context of improving supply and demand matching on electrical grids, improved energy efficiency can have substantial effects on the renewables and fossil fuel resources needed. Importantly, different load end-uses – such as heating, cooling, and lighting – have widely differing temporal consumption patterns. Thus, improvements in energy efficiency in different end-uses have differing effects on the supply mix. Further, as the supply mix changes, the ramifications of improved efficiency of different end-uses also changes. Peaks in the natural gas resource on a renewables-heavy grid are no longer coupled to peaks in demand, but rather coupled to shortfalls in renewable supply. This critical change calls for rethinking the role of energy efficiency in renewables-heavy grids.

To quantify this change, we obtain hourly consumption data from the California Commercial End Use Survey (CEUS) [21] for each of thirteen different end-use categories (*e.g.*, exterior lighting, interior lighting, heating, cooling, ventilation, *etc.*). Scaling these data to match the overall magnitude of commercial consumption for the CAISO generation data in California, we obtain estimated hourly traces for each end-use, shown in Figure 4.12.

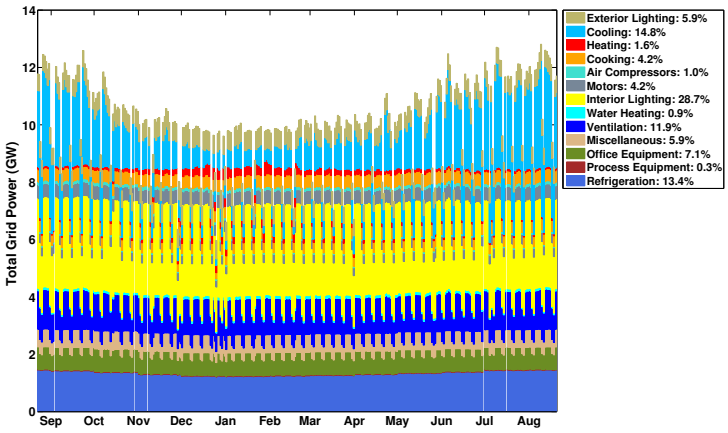


Figure 4.12: CAISO/CEUS - California. Timeseries of end-use consumption over a year for commercial buildings in California scaled to match the overall magnitude of commercial electricity consumption in the state. The data were generated from models for the California Commercial End-Use Survey (CEUS) [21].

We note that these data are not measured, but rather modeled, derived from survey data and appliance models – more details are available in the full CEUS report [21]. Just as in the supply mix, different temporal patterns dominate various categories. Though nearly all classes have significant variations on the daily and weekly scales, three categories – exterior lighting, cooling, and heating – have noticeable variations seasonally.

To assess the efficacy of efficiency measures, we examine two metrics for each end-use on the present-day CAISO grid: cross-correlation with the natural gas generation timeseries and the proportional consumption of the load during peak natural gas hours. The purpose of the former is to measure the similarity of the patterns of end-use consumption and natural gas generation. It stands to reason that efficiency improvements on end-uses that have highly similar consumption patterns to those of natural gas generation (*i.e.*, high cross-correlation) would yield the best results from a natural gas peak reduction perspective. The second metric assesses how "top-heavy" the end-use is relative to natural gas generation. Rather than temporal patterns, this metric aims to capture the proportion of end-use demand at the top of the natural gas generation duration curve.

Figures 4.13, 4.14, and 4.15 show scatterplots of end-uses evaluated on both of these metrics in the California grid. The size of each datapoint represents the mean power consumption of that end-use. For the proportional metric, we use the total proportion of energy consumed by each end-use that occurs during the top quartile of hours by natural gas consumption. For an end-use whose consumption is evenly distributed over all hours, this value should be 25%; greater than that implies a relatively more attractive target for improved efficiency. The difference in the plots is the renewables penetration of the underlying grid represented; for scaling, we use the same methodology described in Section 3.1 with baseload

generation consisting of nuclear and hydroelectric generation ($\mathfrak{B} = 28\%$).

In the present-day grid (Figure 4.13), the three seasonally-dominated end-uses – cooling, exterior lighting, and heating – are the only outliers, while all other end-uses form a relatively tight cluster. Within this cluster, refrigeration, which has the most stable average load, has the highest cross-correlation with natural gas generation, as natural gas is often used as baseload generation in this configuration. Further, the entire cluster has greater than 25% of its consumption during the peak hours; consumption is slightly higher during peak hours. This holds true in the case of cooling, but even more significantly – greater than 35% of energy consumption tends to occur during peak natural gas hours, *i.e.*, during the hot summer months. For heating and exterior lighting, a larger proportion of consumption tends to occur at times other than the peak natural gas hours, *i.e.*, during unconstrained hours where baseload generation is readily available. This explains why cooling is often the focus of existing DR efforts.

As renewables penetration increases to 30% (Figure 4.14), very little changes on this chart. This is because the additional renewables have largely yet to displace the need to operate natural gas facilities, they have only reduced the required capacity of those facilities. However, as renewables increase further to 60% (Figure 4.15), we see a massive change: cooling has become both far less correlated with natural gas generation and far less concentrated during peak natural gas hours. This is because of the large seasonal shift in availability of renewables in deep penetration scenarios; with summer no longer constrained, energy for end-uses like cooling is plentiful and as a result, efficiency effort in that space achieves the least reduction in required natural gas generation. On the other hand, end-uses like heating and exterior lighting, which tend to occur during nights and in the late autumn and winter, now have much higher concentration during peak hours. This finding suggests that efforts at improving energy efficiency of loads should also change as the grid evolves to incorporate more renewables; measures like improving lighting efficiency and encouraging better building insulation will pay more dividends than more efficient air conditioning and cooling systems.

4.5 Comparison to Related Work

The methods in this chapter are related to significant bodies of work in renewables integration, hydroelectric management, energy storage, demand-side management, and energy efficiency. Here, we provide a general overview of concepts in these fields and how they relate to our work.

The plunging costs of wind generation coupled with the challenging unpredictability of wind availability has spawned a rich body of work on renewables integration, containing elements of power systems [41, 127], market operations [15], and load adaptation [99], among others. Though this work focuses on wind integration, many of the techniques and observations are generally applicable to all non-dispatchable renewables. Looking more deeply at the role hydroelectric generation can play in integrating renewables, Dozier provides a good

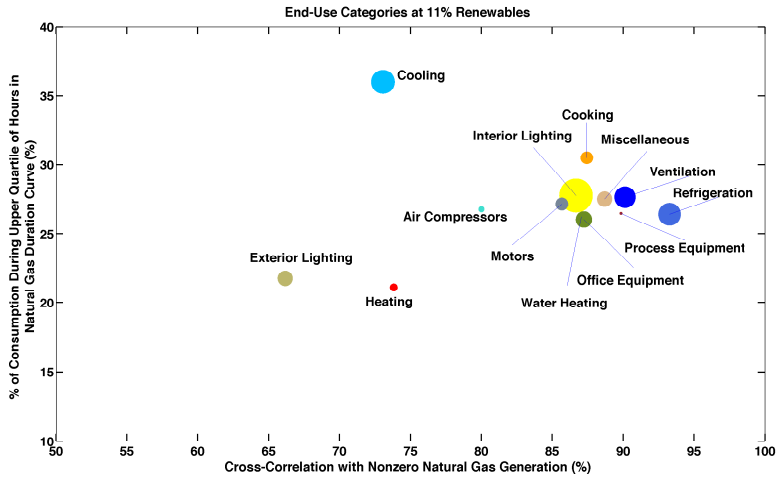


Figure 4.13: CAISO/CEUS - California. Proportion of of end-use consumption in upper quartile of natural gas generation hours vs. cross correlation of end-use consumption timeseries with natural gas generation timeseries on the California grid with 11% renewables penetration (present-day).

overview of the tools used to co-optimize water and power systems [46]. The author discusses issues around using networks of reservoirs to provide firming capacity to renewables.

Energy storage is an area rich with research, but with limited grid-scale deployments. NREL provides an excellent overview of the value of storage on electricity grids, but admits there is further work necessary to understand the role of storage on grids with deep renewables penetration [44]. Others have looked at algorithms for managing storage resources [72, 118] and the value of storage at different timescales [13]. Our work is closest in spirit to the latter, making assumptions of perfect round-trip efficiency and attempting to ascribe value to an ideal storage medium relative to other methods of integrating renewables.

There is a cornucopia of research on adapting loads to become flexible consumers of power, including work in data centers [4, 75, 82], electric vehicles [17], and home appliances [48, 59, 70, 85, 114], among others. Further, substantial attention has been devoted to signaling protocols to achieve certain kinds of demand response [12]. These efforts represent instances of the supply-following loads we implement and assess in this work, and serve as inspiration for their control schemes and methods of evaluation.

In the area of energy efficiency, there has long been argument about the Jevons paradox, which observes that increased efficiency in use of a resource tends to counterintuitively *increase* aggregate consumption of that resource [125]. As a result of this, some argue that the result of energy efficiency standards are overstated [71], while others disagree [83]. Nonetheless, there are enormous industrial and academic efforts in energy efficiency of a variety of energy loads. Though the potential for a 'rebound effect' due to the Jevons paradox is plausible, we believe that targeted energy efficiency is an important piece of the puzzle necessary to achieve deep renewables penetration.

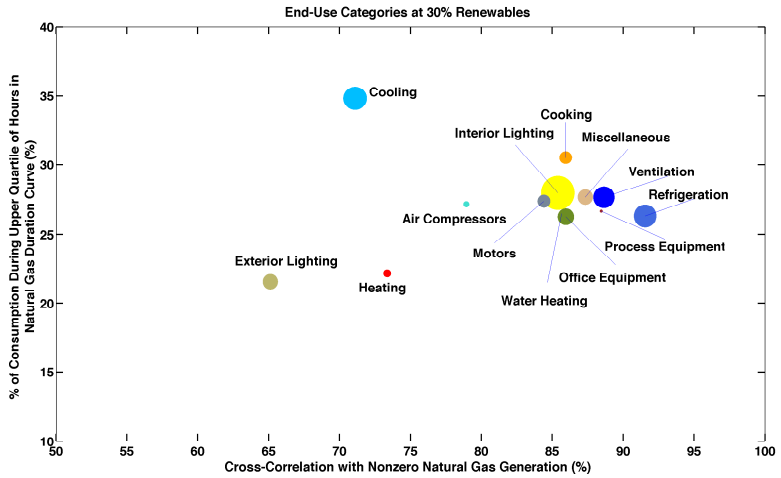


Figure 4.14: CAISO/CEUS - California. Proportion of of end-use consumption in upper quartile of natural gas generation hours vs. cross correlation of end-use consumption timeseries with natural gas generation timeseries on the California grid with 30% renewables penetration.

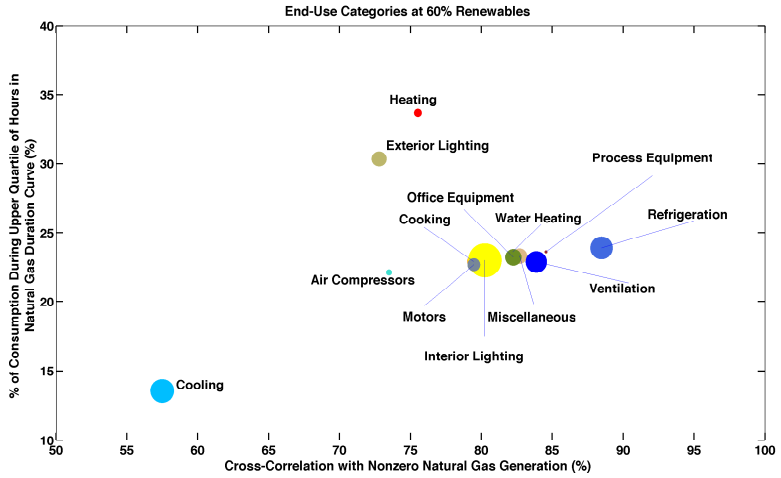


Figure 4.15: CAISO/CEUS - California. Proportion of of end-use consumption in upper quartile of natural gas generation hours vs. cross correlation of end-use consumption timeseries with natural gas generation timeseries on the California grid with 60% renewables penetration.

Chapter 5

Supply-Following Loads

In the previous chapter, we examined a series of techniques for improving the match between supply and demand in a grid with deep penetration of renewables. Among these techniques, portfolio management and grid-scale storage are well-studied and well understood topics. Supply-following loads, on the other hand, are less studied and possess enormous potential in aggregate – large reductions in natural gas generation beyond that possible via changes in how dispatchable supplies are managed and massive investments in grid storage at scales heretofore unrealized. Further, the utility of supply-following loads for improving the match between supply and demand becomes even more substantial as electrical grid generation become more renewable. In this chapter, we push to realize the potential of these loads by implementing and evaluating a set of supply-following loads. We focus on loads that gain their flexibility via either thermostatic or thermochemical (battery) energy storage and employ both simple, threshold-based and more complex, model predictive controls. We adapt existing loads to become supply-following, but our techniques are applicable for creating new supply-following loads as well.

5.1 Supply Variability and Slack

To dynamically match supply and demand down to the appliance-level, we propose *deep demand response* – a distributed conjoining of energy information with physical control systems enabled by pervasive sensor/actuator networks – that goes beyond existing demand response programs [12]. The two keys to matching supply and demand in real-time are predicting the output of renewable sources and controlling the consumption of loads in response to these dynamic predictions. The challenge lies in meeting the quality-of-service requirements of the load while adapting to variations in the source. To model source variability, we consider the CDF of the source *ramp rate*, or change in output. To reason about load adaptability, we use the notion of energy *slack*. Intuitively, slack refers to the amount of time an energy-consuming operation can be advanced or delayed.

To make our proposal more concrete, consider an example of wind as the renewable

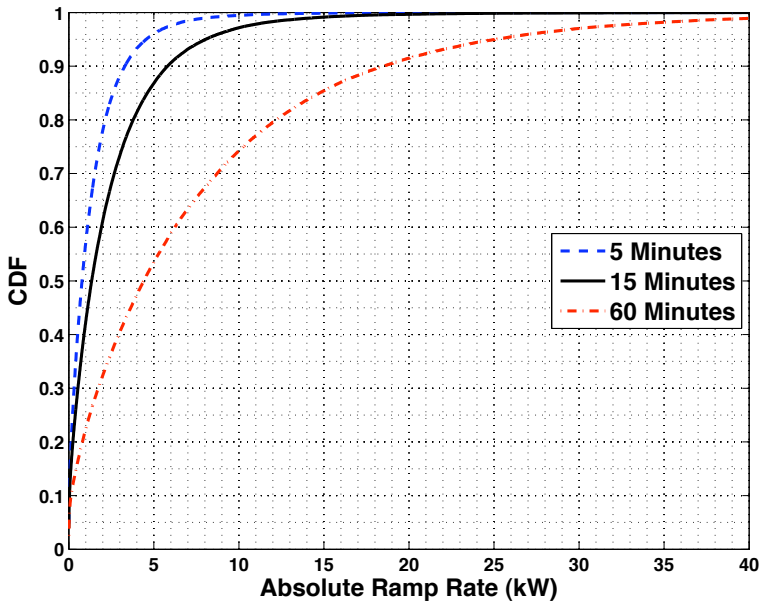


Figure 5.1: Wind ramping. Although wind is fairly unpredictable over longer timeframes like hours and days, its short-term variability, over tens of minutes, is quite predictable. This energy information can be used to dispatch loads and take advantage of slack, either by advancing or delaying energy consumption.

source and a refrigerator as a programmable load. Figure 5.1 shows the variability of a wind source over different time frames. For example, the output power changes by less than 5 kW over a 5 min window 95% of the time, but up to 25 kW over a 60 min window 95% of the time. Now, imagine if this kind of energy information were available to loads (perhaps in aggregate). Then, intelligent, networked loads could adapt their behavior to be greedy or miserly with the energy consumption. For example, a refrigerator could advance the start time of its cooling cycle to consume excess energy, or a temperature setpoint could be slightly increased to reduce energy consumption. In the first case, pre-cooling changes the phase of subsequent cooling cycles, while in the second case, changing the temperature setpoint (for a cycle or two) reduces the energy consumption.

To represent the range of dispatchability opportunities – varying from low- to high-power loads, short- to long-running operations, and one-time to continuous tasks – we introduce the notion of *slack*, or the potential of an energy load to be advanced or deferred without affecting earlier or later operations or outcomes. In critical path analysis, slack refers to the scheduling flexibility in a non-critical path task that keeps the task off the critical path. Slack is a basic and well-understood concept in many disciplines, but in this instance, we measure it in common units of energy and apply it to the operation of physical systems, where the goal is not completion time. Usually, in physical systems, we are concerned with some other input and output variables – in the case of a kitchen fridge, we might ask: *How*

sculptable is the refrigerator load?

More important is the question of how much can we shift the refrigerator's compressor cycle (and the resulting power draw) while keeping the temperature within an acceptable operating envelope. To make the discussion more concrete, consider Figure 5.2(a), which shows one fridge operating cycle consisting of forced cooling (from time $t_0 = 00{:}00$ to time $t_1 = 00{:}11$) and natural warming (from time t_1 to time $t_2 = 00{:}59$). The lower temperature threshold, T_ℓ, is at 2.6°C and the upper threshold, T_u, is at 3.4°C. The corresponding power consumption of the fridge is shown in Figure 5.2(b).

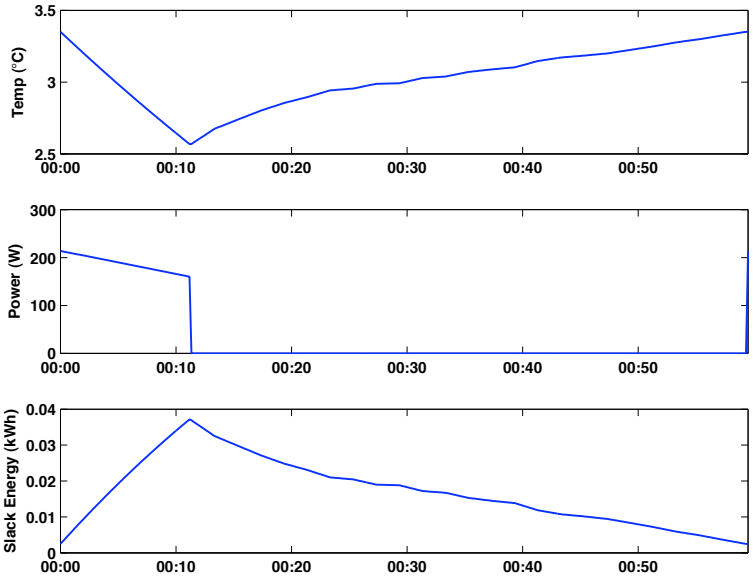


Figure 5.2: The operating cycle of a fridge: (a) temperature, (b) power consumption , and (c) energy slack.

At time t_0, the beginning of the forced cooling cycle, the slack s_w begins at 0. In other words, we can not postpone the forced cooling phase any more, as the temperature has risen to cross the upper threshold T_u. Indeed, this is precisely the control law that a refrigerator follows today. As the time progresses from t_0 to t_1, the energy slack s_c increases according to the accumulation of the energy input $E(t)$ into the system as a function of the system power $P(t)$; in this case this is the power consumption of the compressor.

$$s_c(t_c) = E(t) = \int_{t_0}^{t_c} P(t)\, dt$$

For simplicity, we assume that the conversion of electrical energy to cooling energy by the compressor is perfect – though not true, we believe it is unnecessary to differentiate between slack energy and energy input for this analysis.

Now let us consider the slack during the natural warming phase. Initially, the slack is at its peak value for this cycle, as the cooling phase has just ended and the temperature

has dropped to its lower threshold T_ℓ. As the warming phase starts, the temperature begins to rise and the slack energy stored in the system degrades. To determine the slack during the warming phase, we must find the slack during the cooling phase with the corresponding temperature in the warming phase. Let $T_w(t)$ be the temperature during the warming phase at time t, where $t_1 \leq t \leq t_2$ and let $t_c(T)$ to be the time during the cooling phase as a function of the temperature T. Recall that the slack in the cooling phase is $s_c(t) = \int_{t_0}^{t} P(t)\,dt$. We can now substitute $t_c(T_w(t))$ for t, which gives the slack s_w during the warming phase as

$$s_w(t_w) = \int_{t_0}^{t_c(T_w(t))} P(t)\,dt$$

where $t_1 \leq t \leq t_2$.

Computing the slack using collected values for P(t) and T(t), we can calculate the slack energy as a function of time over a cycle of the refrigerator, as seen in Figure 5.2(c).

Assuming $t_w(T)$ and $T_c(t)$ are known or can be computed, we can plot slack as a function of time, as Figure 5.2(c) shows. More simplistically, energy slack closely follows a banking metaphor – as energy is used by the system to do work (in this case, as the compressor operates), slack is accrued in the slack "account." Then, during the natural warming cycle, the account is drained and the slack is slowly reduced. Also, just as deposits can remain in a bank account, slack can remain in a system across cycles in the form of stored energy. In the case of a battery, this is the capability of the battery to introduce energy into the system later. For a refrigerator, this is the thermal energy stored in the contents of the fridge, whether it is food, liquids, or even air (though these each have very different capacities for energy storage, as we discuss later in Section 5.2.1). The ability to quantify, predict, and utilize slack to achieve a flexible energy load coupled with awareness of supply variability enables supply-following capability. Supply-following is made possible by the capability to both delay energy consumption until a later occasion as well as advance future energy consumption and consume more power now than otherwise planned.

5.2 Thermostatic Loads with Simple Controls

There is an enormous diversity of electrical loads on the grid today, many of which have some flexibility in their operation. In this section, we discuss two broad categories of these loads: thermostatically-controlled loads [73] and scheduled with slide loads. To understand and ultimately employ the slack inherent in the operation of these loads, we construct models of operation and use these models to predict operation in different conditions. We use this technique to create models of thermostatic loads – a refrigerator and a residential heater – which we then control using a simple thresholding technique. This process shows a basic instantiation of a supply-following load, and lays the groundwork for the more complex creations that follow.

5.2.1 Modeling Loads

Having developed a metric for demand dispatchability, we now describe the creation of models of individual consumers of energy with an eye towards building an integrated model of energy loads found in a building. Here, we examine thermostatically-controlled loads – namely, a refrigerator and a heater – as well as loads that are scheduled with slide – a clothes washer, clothes dryer, and coffeemaker. These models can then be composed and sculpted in response to energy generation variations.

5.2.1.1 Thermostatically-Controlled Loads

A common class of energy loads seek to maintain the temperature of a medium within predetermined constraints; the targeted temperature is called the *setpoint* and the range of constraints is called the *guard band*. In most cases, this is done through closed-loop feedback, actuating when a temperature sensor detects that the lower or upper bound of the guardband is reached. The medium under control varies – sometimes liquid (e.g. water in a water heater), solid (e.g. food in a refrigerator), or gas (e.g. air in a building) – as does the duty cycle of the actuation operation for maintaining the desired temperature range.

In this section, we demonstrate how to generate a model of an example thermostatically-controlled system, with the purpose of enabling accurate simulation of operation under different constraints. Using the empirical vehicle of a refrigerator, a characterization of the warming and cooling phases of the system is performed to generate a distribution of curves to be used in the model. Then, the curves are followed and actuation is performed according to preset control laws to simulate operating behavior of the system. We then show how this basic technique can be applied to another thermostatic system, the heating system in a house.

The experimental data gathered for this study was collected using a network of four wireless devices, each with a Sensirion SHT15 temperature and relative humidity sensor and a Hamamatsu S1087 photodiode, and 1 wireless AC Electricity Meter [68]. Three of the climate sensor suites are deployed inside the refrigerator, attached to the underside of different shelves, while the fourth climate sensor suite is attached to the outside of the fridge. The AC Electricity Meter is in series with the refrigerator/freezer power connection. The refrigerator used for the measurements is an 18 cu. ft. General Electric model GTS18FBSARWW. The climate sensor suites are sampled every 10 seconds, while the electricity sensor is sampled every second – this helps for capturing transient electrical loads due to compressor start-up.

Each sensor or sensor suite is attached to a low-power, wireless Telos mote [101] that is running an IPv6-compatible networking layer. The motes form an ad-hoc network along with a laptop that acts as both an edge router for Internet access and a data storage entity, recording data samples into a MySQL database. Having all of the sensors on a network was important for correlating events between sensors as well as automating the gathering of sensor data.

The fridge was monitored for over 3 months at various set points, though only six days of

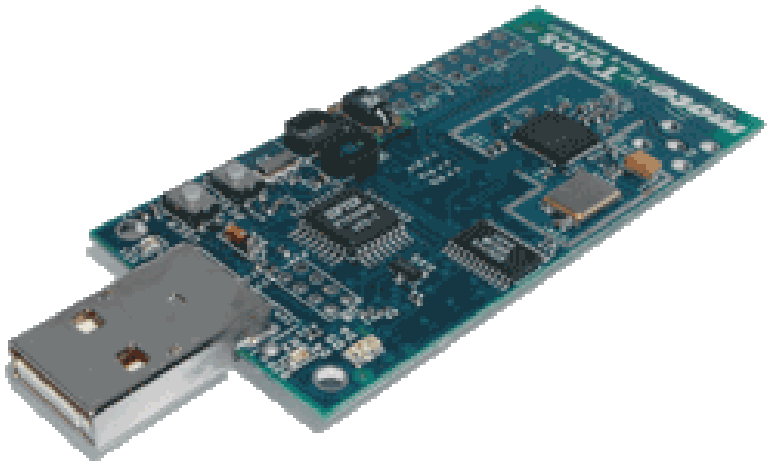


Figure 5.3: A Telos mote with a Sensirion SHT15 temperature and relative humidity sensor and two Hamamatsu photodiodes, used for monitoring refrigerator environmental conditions.

data, all at the same set point, were necessary for this study. Over the six days, each cycle of the refrigerator was identified by observing electricity consumption – if the fridge was consuming electricity, it was in the cooling phase, otherwise it was in the warming phase. Over six days, this identified a total of 153 cycles, or just over one cycle per hour. Then, data from the photodiode were used to identify any cycles that may have been perturbed by a door opening event, as this fridge is in regular use. These cycles, numbering 55 out of 153, were removed from the analysis so that the effects of human perturbations on our model system are minimized.

To model the warming phase of the temperature curve, a number of different techniques were used. Looking at Figure 5.2(a), the curve resembles a decaying exponential approaching an asymptote. Physically, this squares with the intuition that the air temperature in the refrigerator asymptotically approaches the outside air temperature. The rate that this occurs is governed by the constant resistance of the fridge walls to emitting cool air as well as the diminishing temperature difference between the inside and outside air, according to Newton's Law of Cooling. Concurrently, the thermal mass of the contents of the fridge changes temperature more slowly, and thus counteracts the warming of the air temperature. In theory, modeling this temperature response calls for a regression using two exponential terms. In practice, however, the portion of the curve captured (roughly between 2.5 and 3.5°C) is a small portion of the entire warming curve, which would approach an outside air temperature of approximately 19-21°C. Thus, with a limited set of points, curve fitting solutions that minimize mean-squared error generally do not capture the shape of the curve in the limited domain, resulting in fits that are essentially linear over the domain under consideration. Thus, in order to accurately fit the warming curves empirically observed, we use a linear spline model that interpolates between the observed values and extrapolates to

a wider domain by extending lines at the same slope as the end of the observed domain. We believe that the high rate of change of the temperature early in the warming phase is as a result of the low heat capacity of the air inside the fridge, which changes its temperature quickly. Once the air is saturated, the fridge temperature change transitions towards being linear, and this pattern continues outside the measured domain. Thus, our linear spline extrapolation fits the behavior of the warming curves accurately for both the measured domain as well as the extended domain.

For modeling the cooling phase of the temperature curve, the factors contributing to the warming curve remain relevant. However, another component, the forced cooling action of the refrigerator compressor, not only counteracts the warming, but also cools the fridge at a significantly faster rate. This is evident in the durations of the cooling phase and warming phase – 12.3 minutes and 45.3 minutes over the 98 unperturbed curves, respectively. Thus, the extra power of the compressor allows the fridge to cool nearly four times as fast as it naturally warms. In practice, this diminishes the effects of the natural warming components in designing a model of the cooling phase of the fridge. In addition, the power consumed also follows an exponentially diminishing curve during each cooling phase. For both, we use a simple exponential decay model with three parameters, according to the equation

$$T(t) = a_1 - a_2 * (1 - e^{(a_3 * - t)})$$

For each run of the fridge simulation, a set consisting of a warming, cooling, and energy curve is selected. These curves are not chosen independently because the shape and cooling rate of the cooling curve is dependent on the warming curve it follows, and the warming curves exhibit significant variations over the six days of data. The same holds true for the energy curve – its shape is linked to its corresponding cooling curve. In order to select a set of curves, we create a distribution of the curves using the average warming rate ($\frac{T_{end} - T_{begin}}{t_{end} - t_{begin}}$).

To run a simulation of the refrigerator, an initial condition is randomly chosen from the range of temperatures within the constraints of the operating range (known as the guardband), and an initial state (warming or cooling) is chosen according to the proportion of time spent in each phase (≈20% cooling, ≈80% warming). From there, temperature curves can be plotted according to the phase of the operating cycle. Additionally, values for the power curve can also be plotted concurrently. Whenever the simulated fridge reaches a guardband boundary, the phase is changed and a different curve is used. This simple fridge model, oblivious to its energy input and any other external signals, aims to recreate the behavior of an unperturbed fridge. Without variations in the guardband and no perturbation, the fridge merely repeats the temperature and power cycles seen in Figure 5.2, resulting in a compressor duty cycle of 19.0% with the set of curves selected, though the duty cycle varies between 13% and 35% for other curves.

Having modeled a specific thermostatic device, we would like to emphasize that the technique for constructing a model is generalizable to a wider set of energy loads. To this end, a data trace was collected for the temperature of a house. The house under observation resides in a relatively mild climate where outside air temperature is generally 5-10°C lower

than the house temperature, and contains a heater but no air conditioner. In accordance with this temperature trace, a complementary electricity trace for a heater is synthesized. Note that the magnitude of power consumption is sized for only the fan of the furnace – gas heat is assumed. If an electric furnace was assumed, consumption would be significantly higher.

An important variation between the fridge and thermostat models created is that the fridge operating cycle is forced cooling and natural warming, while the thermostat is opposite – forced warming by the heater and natural cooling as the inside air approaches the temperature of the cooler outside air. Also, in this case, we do not create a distribution of curves, instead simply using a single warming and cooling curve as a proof-of-concept. The resulting trace of the temperature, power, and slack behavior of an oblivious thermostat over three hours is provided in Figure 5.4.

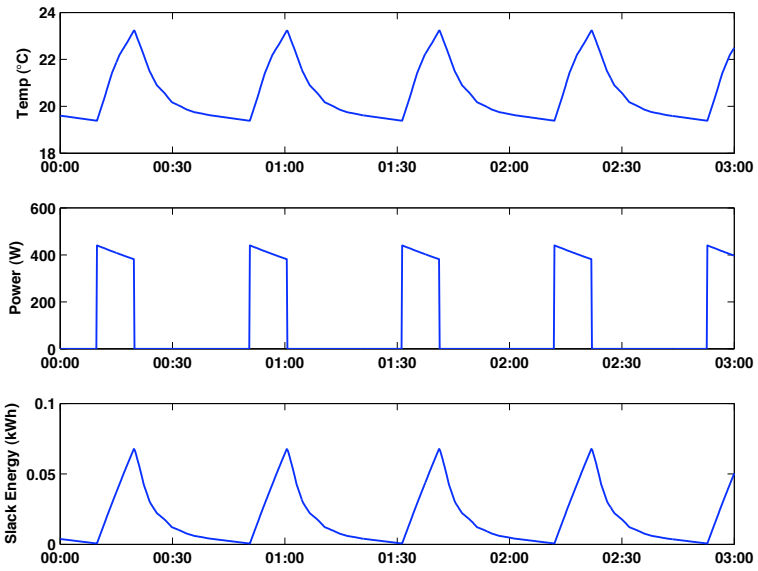


Figure 5.4: The operating cycle of a house thermostat: (a) temperature, (b) power consumption , and (c) energy slack. The thermostat operates in a reverse pattern from the fridge, with an operating cycle that consists of forced heating followed by natural cooling. Thus, the energy slack is phase-shifted by 180 degrees from the refrigerator. The duty cycle of the heater is about 24.6%.

5.2.1.2 Scheduled with Slide Loads

The other class of dispatchable energy loads we examine in this work are those that can be scheduled for completion in some timeframe. We expect that the flexibility of this type of load will depend highly on the quality-of-service demanded by the user; if people are not

incentivized to be patient for completing their operations (i.e. if there is no preferable tariff or other benefit given for delaying the operation), then flexibility will be limited.

We recognize that typical appliances do not yet provide this capability. However, we believe that the emergence of plug-in electric vehicles – each of whose power consumption roughly equals that of an entire house, essentially requiring some sort of smart charging for a stable electricity grid – will make scheduled appliance operations more familiar to the general public.

Device Type	Model	Energy Per Operation	Arrival Schedule
Clothes Washer	Whirlpool Imperial	0.145 kWh	One load per day, starting randomly between 6 AM-9 PM
Clothes Dryer (Gas)	Whirlpool Imperial	0.199 kWh	One load per day, starting randomly within 1 hr. after washer completes
Coffeemaker	Cuisinart 12 Cup	0.249 kWh	Daily at 4 AM, remain warm until 7 AM

Table 5.1: Loads that are scheduled with slide. Each has been modeled from empirical data; job scheduling is assumed for a family of four.

Generating models for these appliances is straightforward: traces of example appliances were obtained from the Lawrence Berkeley National Labs Appliance Energy Use Data repository [79]. Each appliance was assigned an arrival pattern according to expected usage. Table 5.1 shows these arrival rates, and Figure 5.5 shows traces of a clothes washing machine, clothes drying machine, and coffeemaker. Note that the clothes dryer is in fact a gas model; using an all-electric clothes dryer would significantly increase electricity consumption.

5.2.2 Supply-Following Thermostatic Loads

In this section, we exercise our models of thermostatically-controlled loads to study alternate control strategies for the loads under study. We first explore the parameters and boundaries in load flexibility. As stated earlier, the primary purpose of exploring slack in energy loads is not for energy efficiency, but instead for advancing or deferring operations to better match energy consumption to generation. However, flexibility (and thus, slack) is also improved by softening guardband constraints. Figure 5.6 shows the effect on the heater average power consumption from its baseline average at 266.2 Watts by altering the guardband constraints.

That thermostatically-controlled loads are able to "walk" this graph – consuming comparatively more power at certain times while using less power at other times – allows them to follow source generation. This ability remains true for the fridge as well, and enables a range of flexibility in control strategies. In Figure 5.7, we show three different control scenarios of the house heater, with the rows representing temperature, power consumption, slack energy, and a two-day profile of wind energy from a wind farm in Minnesota. The gray area (lightly shaded) on the wind energy plots represents the maximum amount of "portfolio" or

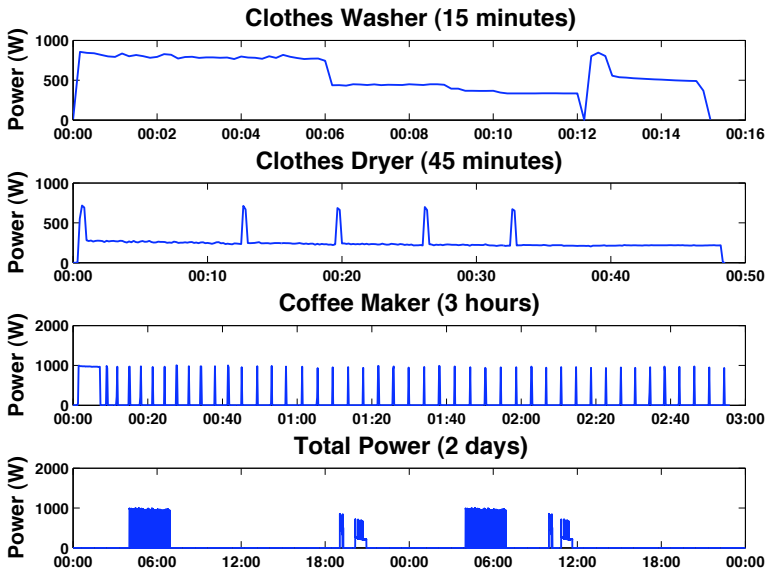


Figure 5.5: Traces of power consumption of a clothes washer, clothes dryer, and coffeemaker. From these traces, a model over two days is created.

non-renewable energy drawn throughout the two-day period. The wind supply generation is the same for all three scenarios.

The first column represents an oblivious heater, which only considers its measured temperature when deciding if the compressor should be actuated. As such, it maintains a tight guardband and consistent slack energy profile. All of the wind plots have been scaled to provide 50% of the total energy needed to power this oblivious heater.

The second column ("supercool") represents a heater that allows the temperature to cool beyond its normal lower bound by 2°C when faced with a renewable energy shortage. The threshold for an energy shortage, represented by the maximum power consumption of the heater when operating, is indicated by the dashed line on the supply power graph in Figure 5.7. Thus, when wind energy is available, the heater operates in its standard tight guardband, consuming more energy.

The third column ("wide guardband") takes this concept further – this heater not only reduces power consumption when faced with a deficit of renewable energy, it also attempts to increase power when faced with a renewable energy surplus by running the heater for longer. In addition, to reduce the discretization of its response, this fridge scales the change in its guardband boundaries by the magnitude of the deficit or surplus of renewable energy. For example, in the late evening of the first day, wind energy far exceeds the threshold, the heater increases its guardband relatively more as compared to the early evening period, when wind energy barely exceeds the threshold. This change results in less energy saved than the "supercool" case because the proportional response is always less than or equal to 2°C.

What these scenarios show is that the heater can reduce the frequency of its cycles, and

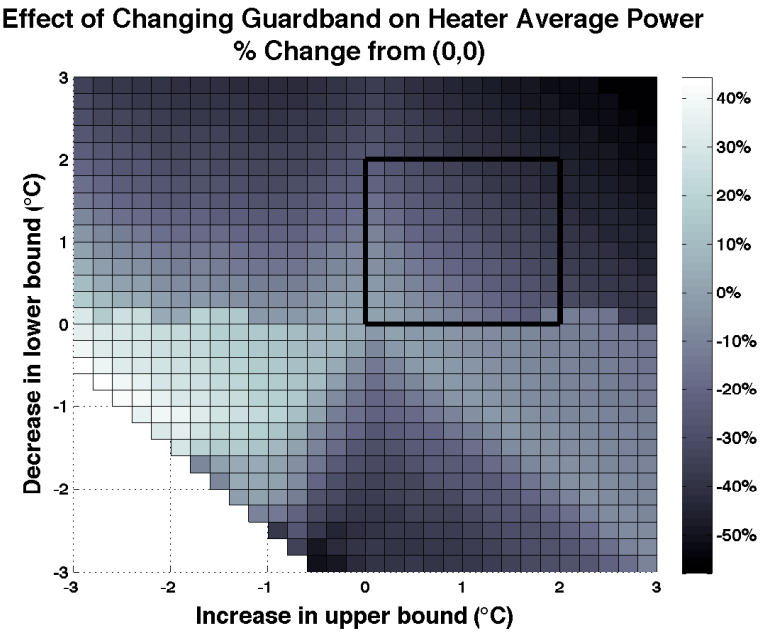


Figure 5.6: The effect on average power consumption from changing the heater guardband constraints. The black square represents the region explored in this work (+/- 2°C). Within this region, power consumption varies by 47%, enabling significant responses to variations from renewable energy supplies. Note that a positive decrease in the lower guardband boundary reduces the lower guardband temperature.

thus its energy consumption, when faced with an energy deficit. A natural follow-up question is whether house occupants are amenable to a wider guardband at certain times – this has been addressed widely in the building comfort and demand response literature. [64]

What is the magnitude of savings in these scenarios? Figure 5.8 compares the three; the first plot shows total energy consumption while the second considers the balance between renewable and portfolio energy. The ability to preferentially loosen guardband constraints saves significant energy – over 33% by reducing only the lower guardband boundary and over 50% while altering both the lower and upper boundaries, even with responses proportional to the energy deficit or surplus. The increase in energy savings from relaxing the lower guardband boundary is not surprising, but the large additional reduction from increasing the upper boundary is. This is because longer than standard actuation cycles (heating phases, in this case) ensure that the heater has a lower overall power consumption than the default scenario. We leave quantifying the limits of the potential reduction to future work.

Looking at the second plot of Figure 5.8, we see that the increase in the proportion of energy consumed that is renewable is modest; at best, renewable energy is 60% of the total proportion versus 50% in the initial scenario. This is a result of the burstiness of heater power consumption – when a single heater is measured, its consumption appears as a series of spikes,

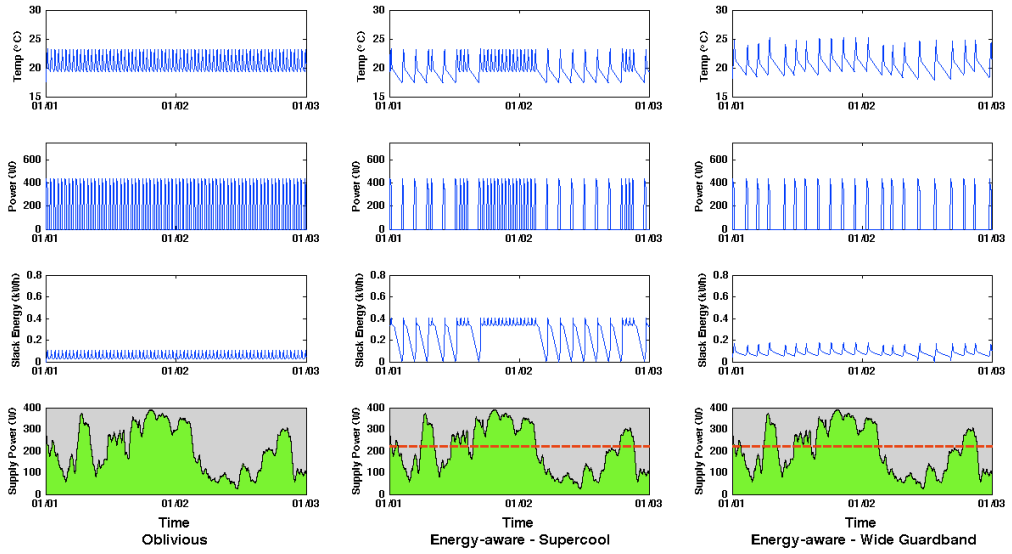


Figure 5.7: Three scenarios of heater operation. The first column represents an "oblivious" heater that operates without any outside information. The second column represents a heater that allows the house to get cooler when faced with an energy deficit. The third column represents a heater that allows the house to get cooler when faced with an energy deficit as well as warmer when presented with an energy surplus.

each of which may overwhelm the existing supply of renewable energy. However, the real benefit of this supply-following strategy becomes apparent when a large number of devices are able to shift their consumption; by randomly distributing consumption, the effects of the burstiness are reduced and the smooth consumption pattern follows the energy availability signal. We look into the effects of this technology in larger populations in Section 6.1.1.

5.3 Thermostatic Loads with Complex Controls

Beyond simple threshold-based controls, models of flexible loads can be used to enable prediction in advanced control systems. By simulating future operation of the load under different system inputs, control decisions can be made to achieve a desired future system output. In this section, we implement and evaluate a thermostatic load that uses an advanced control method called model-predictive control to maximize load flexibility. The particular load under study is a building temperature control system that aims to achieve maximum energy efficiency; this is equivalent to a supply-following load with the available supply always at zero. Given this input, the system tries to reduce its energy consumption as much as possible, and thereby shows the large amount of flexibility possible in this load.

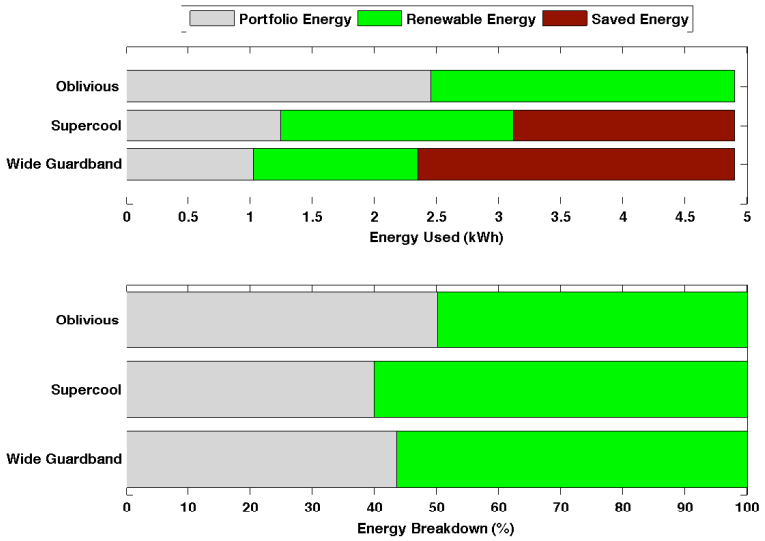


Figure 5.8: The energy breakdown of the three heater scenarios. The top plot emphasizes the energy savings of increasing the guardband, while the bottom plot shows the improvement in the proportion of total energy that is renewable.

5.3.1 Model-Predictive Control on a Residential HVAC System

As thermostatic systems that incorporate large volumes of air as thermal mass, Heating, Ventilation, and Air Conditioning (HVAC) systems in buildings represent ideal candidates to become supply-following loads. With a third of all building energy going towards this end use, and buildings accounting for 73% of the overall electricity consumption in the U.S., the potential magnitude for impact is also large [18]. In particular, a common route towards reducing HVAC energy consumption is the design of new, more efficient equipment. However, buildings and equipment are often slowly replaced [38], leading to interest in retrofitting existing HVAC systems to improve efficiency.

There is a large variety of HVAC equipment, making the problem of retrofitting for efficiency a challenging one. Yet this provides an opportunity for the industrial and academic communities, because different equipment requires different approaches to retrofitting. There has been considerable work on modeling and efficient control of different HVAC equipment [61, 42, 96, 87, 95]; here we focus on modeling and control for a single-stage heat pump air conditioner (AC) that cools a single area. Our goal is to improve its energy usage while keeping the room temperature close to a desired value.

We begin by describing the Berkeley Retrofitted and Inexpensive HVAC Testbed for Energy Efficiency (BRITE), which we built for the purpose of control experimentation *in situ*[1]. The BRITE testbed controls the room temperature of a computer laboratory on

[1]BRITE is the result of a collaboration with Anil Aswani and Neal Master.

the Berkeley campus that is actively used by students. In our setup, we can monitor the room temperature and actuate the AC unit through a networked thermostat. A computer, connected through a local area network (LAN), is able to compute a control action and relay it to the thermostat. This is an *in situ* platform rather than a traditional controlled testbed or simulator, which introduces challenges in how to compare different control strategies.

Next, we examine the electrical power characteristics of the heat pump. Understanding these features is important for saving energy, which are likely common to many other HVAC systems. We use a model predictive control (MPC) technique, which allows these characteristics to be directly encoded into our control scheme. This enables our designed control scheme to make a tradeoff between minimizing transient and steady-state energy consumption and keeping the temperature within a comfortable range.

Before we design the MPC scheme, we must identify a model for the temperature dynamics of the room. Room occupancy contributes significantly to this, but it is difficult to model because it is highly nonlinear with respect to time. One approach has been to model this by combining occupancy sensor measurements with models of human behavior [80]. Here, we use a semi-parametric regression approach [14, 106] that can estimate the heat load due to occupancy directly using only temperature measurements. The advantage of this is that we do not require extra sensors beyond the temperature sensor already in the thermostat.

We specifically make use of a provably safe and robust learning-based MPC technique [8] that can be used to estimate occupancy from only temperature measurements and compensate for it within the control action. A typical thermostat uses two-position control [62]. In two-position control, the user selects a target temperature called the setpoint and an allowable range around the setpoint called the guardband. Our experiments show that learning-based MPC reduces energy consumption by 30-70% compared to two-position control. We conclude by analyzing the transient and steady state behavior of the AC to understand the causes for this improvement.

5.3.1.1 Berkeley Retrofitted and Inexpensive HVAC Testbed for Energy Efficiency

The Berkeley Retrofitted and Inexpensive HVAC Testbed for Energy Efficiency (BRITE) is a system for testing different control strategies on an AC unit that cools a computer laboratory on the Berkeley campus, as shown in Figure 5.9. It is built using commodity parts, though the computers could be replaced with microcontrollers. The strength of this structure is that it scales to building-wide systems. Furthermore, the convexity of our MPC schemes makes them computationally scalable, while adapting to varying occupancy.

In this testbed, the LoCal server gathers and stores sensor data in a Simple Measurement and Actuation Profile (sMAP) database [39]. A control computer accesses the Internet and LoCal server to get weather forecasts and sensor data, and runs an MPC scheme that computes a control input that is sent through the LoCal server to the thermostat. The thermostat converts this input into an electrical signal that is transmitted to the AC.

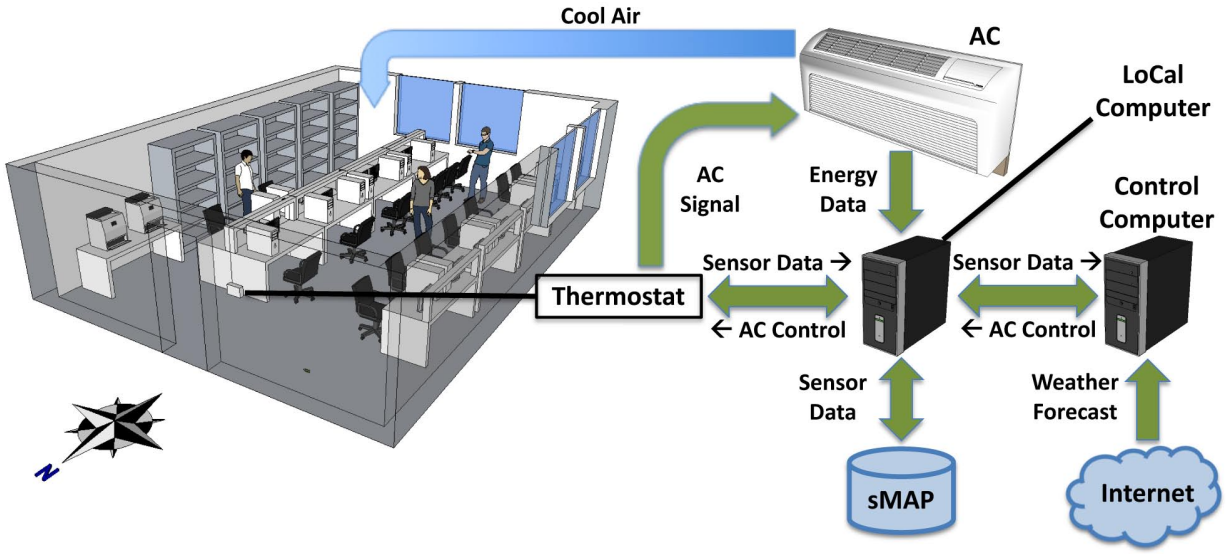


Figure 5.9: The Berkeley Retrofitted and Inexpensive HVAC Testbed for Energy Efficiency (BRITE) allows testing of different control strategies for controlling an AC in order to explore tradeoffs between energy consumption and tracking a temperature setpoint.

The Berkeley LoCal project aims to produce a network architecture for localized electrical energy reduction, generation, and sharing by examining how pervasive information can fundamentally change the nature of these processes [103]. A key component of this is the use of sMAP [39] to exchange physical data about the systems involved. This allows producers of physical information to directly publish their data in a format for consumption by a diverse set of clients. We use temperature measurements from a networked thermostat in BRITE, though we also have wireless plug-load power consumption [67] and temperature sensors. The ability to easily integrate streams of sensor data is critical to the scalability of BRITE to entire buildings.

The BRITE testbed is deployed in a student computing laboratory on the ground floor of a large engineering building; a CAD model of this laboratory is shown in Figure 5.9. The room is 640 square feet and has external windows on its south and west walls. The lab contains 16 desktop computer workstations along with two laser printers. Occupancy of the room peaks at over 20 occupants and varies throughout the day, week, and semester based principally on the dates of projects, assignments, and exams.

The HVAC system is a single-stage electric heat pump, meaning it operates in two states (*i.e.*, on or off). One important reason this specific laboratory was chosen for the BRITE testbed is that it has a dedicated HVAC system (*i.e.*, not shared with any other spaces). This allows us to test the cooling independent of other rooms.

A Proliphix brand NT160e model thermostat controls the AC. It is a modern thermostat that uses Power over Ethernet (PoE) for its power and networking functionality. This ther-

mostat has an application programming interface that allows computers on a shared network to communicate with and control it. We can transmit commands to the thermostat and also receive diagnostic information on the thermostat settings, current HVAC state, and room temperature.

To implement the MPC scheme, we use a dedicated control computer to avoid disrupting the other processes and databases running on the LoCal computer. Note that a production or commercial version of this system could combine the functions of the LoCal computer and the control computer into a single, centralized management system, running either on a server or even an embedded controller.

The control computer runs a 64-bit version of Ubuntu, a Linux operating system. The control loop is implemented in MATLAB, using the SNOPT solver [55] in the TOMLAB library to implement the learning-based MPC scheme [8]. When needed, polytope operations were computed using the MPT toolbox [76]. Additionally, weather forecasts are gathered from the NOAA National Weather Service using a Python script.

The objective of the BRITE testbed is to minimize energy consumption while keeping occupants comfortable. However, quantifying comfort is difficult because it depends on individuals' physiology and psychology. The ANSI/ASHRAE standards [5] are defined in terms of predicted mean vote (PMV), a complicated nonlinear function of indoor air temperature, human activity, relative air velocity, the occupants' clothing, and many other variables that can be difficult to measure or estimate [58].

Alternative metrics exist. The Occupational Safety and Health Administration (OSHA) [111] does not have regulations but provides guidelines of 68–76°F, or approximately 20–24.4°C. Additionally, there are guidelines on temperature deviation. Category A thermal requirements [5] dictate a temperature range of 2°C, while Category B thermal requirements mandate a temperature range of 4°C. The most relaxed requirements are Category C that mandate a range of 6°C.

These alternative metrics of comfort are particularly attractive because they are equivalent to temperature bands, and simplify real implementations of HVAC systems. Based on these, we keep the temperature near the middle of this range (22°C) in the BRITE testbed. Our experiments on the testbed aim to satisfy Category A requirements because these are the strictest and consume the most energy. More specifically, we use Category A as a range preference for the learning-based MPC and Category B as hard constraints on the temperature range. Future directions can consider smart methods for switching between different Category requirements based on, for instance, network-level load and demand signaling or occupancy estimates.

5.3.1.2 Electrical Characteristics and Energy Consumption of a Single-Stage Heat Pump

Heat pumps [56] are very common devices used in HVAC, particularly for homes. A heat pump consumes electrical energy to compress gas. This compressed gas is used to transfer thermal energy from a low temperature region to a high temperature region. In the case of

the BRITE testbed, the heat pump takes thermal energy from a computer lab and transfers it outside.

A single-stage heat pump, like the one used in BRITE, has an on-off control scheme. This means that the heat pump is either doing work at its highest rate (on) or at its lowest rate (off), which leads to particular patterns in its energy usage. Understanding these characteristics is important for the design of energy efficient HVAC systems, allowing them to be explicitly considered and compensated for in an MPC framework.

Figure 5.10 shows experimentally measured data of a typical power consumption profile for the HVAC in BRITE. A striking feature is that there are both transient and steady state behaviors. There is a spike in power consumption immediately after the heat pump is turned on; this transient electrical usage decays after approximately 1 minute, and the power consumption reaches a steady state. Intuitively, the increased transient power consumption is a penalty for turning the heat pump on. Physically, the transient power consumption is due to inrush current drawn by the electric motor in the compressor of the heat pump.

As mentioned earlier, this profile highlights some important issues regarding energy usage. The transient spike in power consumption suggests an objective that minimizes switching. This is not typically considered in an explicit manner, and in fact makes the implementation of a controller in digital hardware difficult unless some approximation is used. Furthermore, the steady state energy usage is linear in the control, matching the cost used in other work [96]. This differs from a commonly used objective function which is quadratic in the duration of control [87].

5.3.1.2.1 Pulse Width Modulation Control The single-state heat pump is strictly speaking a hybrid system [120, 86] because it has two logical modes corresponding to the pump being on or off. Fortunately, we can considerably simplify the design of a controller by considering sampled control. As such, we use MPC to compute a new control action $u[k]$ at intervals of once every 15 minutes. We chose this rate because switching more frequently than once every 10 to 15 minutes can physically damage the heat pump. Pulse-width modulation (PWM) is used to convert the discrete time control $u[k]$ into a continuous signal that turns the AC on or off [19]. For this reason, the value $u[k]$ can also be interpreted as a duty cycle.

We set the constraints on the input for the MPC to $u[k] \in [0, 0.5]$. The reason for the choice of 0.5 as an upper value is because the thermostat does not stop cooling the room when it is turned off. This is discussed in more detail in Section 5.3.1.3, but the choice of 0.5 ensures that the control action at one discrete time sample does not affect the control action at the next time sample.

5.3.1.2.2 Measuring the Electrical Energy Consumption in BRITE It is important to be able to compute the energy consumed by the AC in the BRITE platform, given the input that the AC receives. An estimate of the energy consumption is used in the cost function of MPC; this estimate is useful for being able to compare different control schemes. To be able to provide this equation, we need to make a few definitions. Define the vector

$\mathbf{u}_m = \big(u[m] \quad \dots \quad u[m+N-1]\big)$. The term $\|\mathbf{u}_m\|_0$ counts the number of nonzero entries in the vector $\mathbf{u}_m$. Also, the values r, λ are constants which are used to compute the energy consumption. The value N is the number of discrete time steps (recall that each time step corresponds to 15 minutes) over which the energy consumption is to be computed.

The steady state energy consumption of the AC over $N/4$ hours in units of (kWh) is given by $\sum_{k=0}^{N-1} r/4 \cdot u[m+k]$, where $r = 3.7kW$ is the average rate of steady state energy consumption in the BRITE platform (compare to Figure 5.10) and the value 4 is used to compensate for the fact that $u[m+k]$ is the control for 1/4 of an hour. Furthermore, the AC consumes $\lambda = 0.015kWh$ of energy every time we turn the AC on; this corresponds to the area of the triangle in Figure 5.10 formed by the transient energy. The total energy used over N time steps is given by

$$E_{actual} = \sum_{k=0}^{N-1} r/4 \cdot u[m+k] + \lambda \|\mathbf{u}_m\|_0. \tag{5.1}$$

Unfortunately, the $\|\mathbf{u}_m\|_0$ term is not convex in $\mathbf{u}_m$. Convexity is important for the computations of the MPC that has to solve an optimization problem at each step. To simplify the computations, we make a standard convex relaxation [32] and replace the term $\|\mathbf{u}_m\|_0$ with $\|\mathbf{u}_m\|_1$. This relaxation is powerful: when it is used in the cost function of an optimization problem, it actually leads to having many of the $u[m+k]$ be equal to exactly zero [32]. In this way, it can reduce switching of the AC.

This approximation yields a convex equation for the energy consumed $\sum_{k=0}^{N-1} r/4 \cdot u[m+k] + \lambda \|\mathbf{u}_m\|_1$. However, we have $u[m+k] \geq 0$, and so we can further simplify the convex cost for energy consumption to

$$E_{convex} = \sum_{k=0}^{N-1} (r/4 + \lambda) \cdot u[m+k]. \tag{5.2}$$

What is surprising about this is that a cost for energy that is linear in the length of control action automatically considers a cost for switching, as long as the inputs $u[m+k]$ are constrained to be non-negative. Stated in another way, this means that a cost that is linear in the duty cycle of the control inherently considers the tradeoff between switching too frequently and the length of the duty cycle.

In practice, (5.1) is used if the actual energy needs to be computed. On the other hand, (5.2) is used if a control action needs to be computed by the MPC. Having these two formulations gives considerable flexibility.

5.3.1.3 System Identification of Cooling Dynamics

An important step towards realizing efficient control schemes for the BRITE testbed is building a mathematical model that describes the impact of weather, occupancy, and AC operation on the temperature of the room. It is important because all MPC schemes inherently require

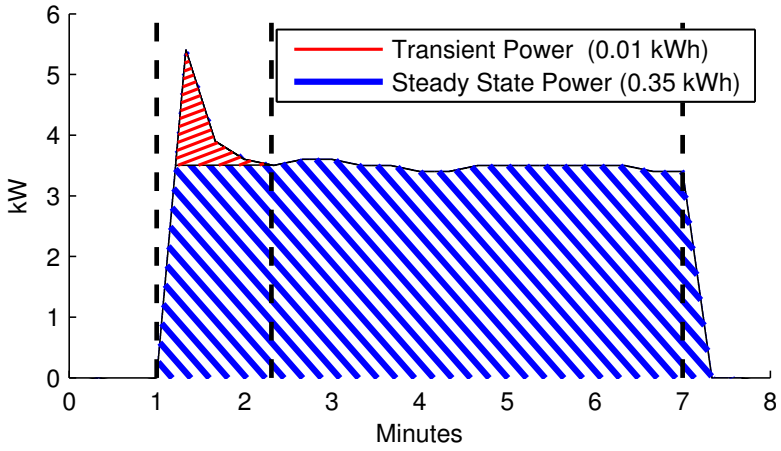


Figure 5.10: Experimental data of a typical power consumption profile during actuation in the BRITE testbed is shown. The first vertical dashed line indicates the time when the heat pump turns on, the second indicates the time when the power reaches steady state, and the third indicates the time when the heat pump turns off. For this particular power profile, the amount of total energy consumed by the transient and steady state is labeled in the legend.

a nominal model in order to be able to optimize system performance. More importantly, identifying a model allows us to estimate the heating load due to occupancy from only temperature measurements. This enables evaluation of the importance of occupancy [11, 87] and techniques that compensate for it.

Though building simulator software [35, 47] models complicated thermodynamic and fluid effects, experimental data from testbeds support the notion that linear models with exogenous inputs [11, 42, 87, 96] can be used to describe rooms where the lighting is not actuated. The main physical effect is convective heat transfer and is described by Newton's law of cooling. This is a linear ordinary differential equation (ODE), and so it may be abstracted as a resistor-capacitor network [42, 87, 96].

5.3.1.3.1 Discrete Time Model There is a "delay" from when the AC is turned off and when it stops cooling the room, due to the dynamics of the heat pump. Specifically, the evaporator which cools the air does not instantly warm up and continues to cool air for some time after the heat pump is turned off [124]. We begin with a discrete time model where each time step is separated by $T_s = 15$ minutes. The advantage of this approach is that the AC behavior gets lumped into a single term that encompasses the modes where the AC is on and then turned off but still cooling. This makes it easier to do the modeling.

With this approach and inspired by the physics of convective heat transfer, we start with the model

$$T[n+1] = k_r T[n] - k_c u[n] + k_w w[n] + q[n], \tag{5.3}$$

where $T[n] \in [15, 30]$ is the temperature of the room in ($^\circ C$), $k_r > 0$ is the time constant of the room, $k_c > 0$ is the change in temperature over 15 minutes in ($^\circ C$) caused by cooling for a duty cycle of $u[n] \in [0, 0.5]$, $k_w > 0$ is the time constant for heat transfer from the room to the outside, $w[n]$ is the outside temperature in ($^\circ C$), and $q[n]$ is change in temperature over 15 minutes due to the heating from occupancy of humans and equipment within the room, as well as other external inputs, in ($^\circ C$). The time constants here are dimensionless.

5.3.1.3.2 Parameter Identification We collected data from 12:00PM to 5:30AM over parts of two weekdays using the BRITE testbed. This portion of a day was used because it exhibits a variety of occupancy levels. The room is actively used by students during the afternoon and evening, with fewer students using the room late at night and early in the morning.

Generally speaking, parameter estimation is usually more accurate when the control inputs are independent of the system states or external inputs (*i.e.*, weather and occupancy). To ensure that this was the case, we actually applied a random input with uniform distribution over $[0, 0.5]$ at each discrete time step; the corresponding PWM control is shown in Figure 5.11(a). Because this only needs to be done once and over a span of about a day, it may be reasonable to allow the temperature in actual implementations to be unregulated for this day. Future work includes designing methods that keep the temperature in a comfortable range while still sufficiently exciting the system.

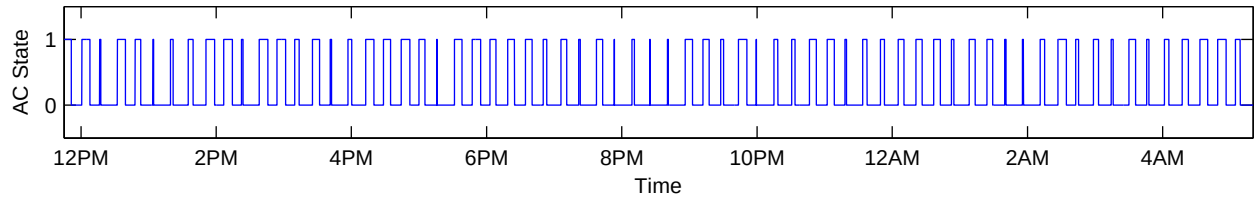
Because we have measurements of T, $w[n]$, and $u[n]$, the model is linear with respect to the parameters k_r, k_c, and k_w. On the other hand, $q[n]$ is not known and is expected to be highly nonlinear with respect to time, because it incorporates heating due to human occupancy and equipment in the room. Consequently, standard linear system identification techniques cannot be used. Identification of models with the form given in (5.3) more generally falls into the class of problems known as semi-parametric regression of partially linear models [14, 106]. An alternative approach is to parametrize $q[n]$, for example with a polynomial or spline basis, and then identify all parameters using nonlinear regression. The difficulty with this is the uncertainty associated with $q[n]$.

Using a technique from Robinson [106], we identified the parameters of the model

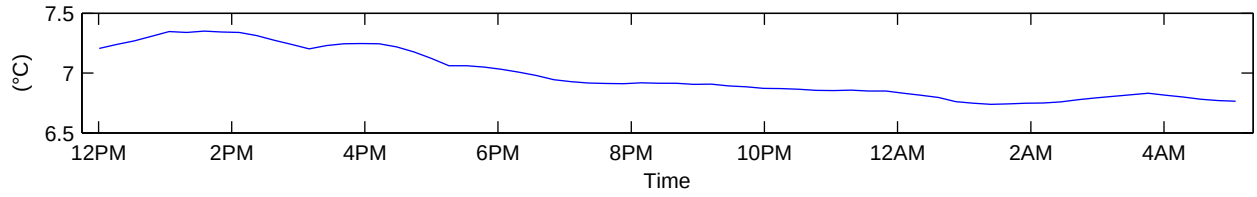
$$T[n+1] = 0.64 \cdot T[n] - 2.64 \cdot u[n] + 0.10 \cdot w[n] + q[n], \tag{5.4}$$

where $q[n]$ is shown in Figure 5.11(b). The experimental room temperature is the solid line in Figure 5.11(c). Similarly, the temperature simulated by the model (5.4) is the dashed line shown in Figure 5.11(c); the initial condition for the simulation was taken from the experimental measurements. Furthermore, the simulation was conducted with the same inputs as were applied to the real BRITE system. The root-mean-squared (RMS) error of the simulation is 0.10 $^\circ C$. The plots show that the model fits reasonably well to the measured temperature data.

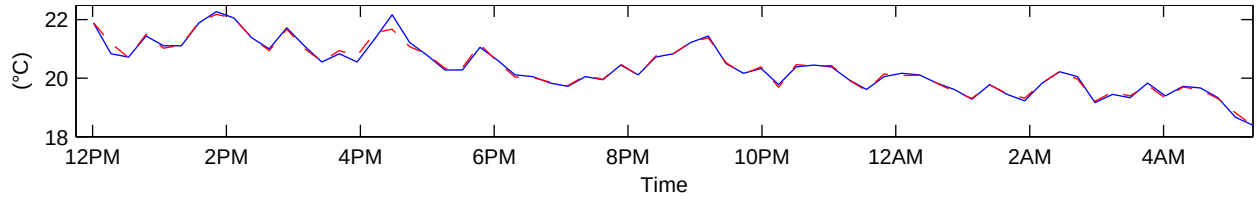
5.3.1.3.3 Impact of Occupancy The identified model (5.4) shows that the role of occupancy is significant in the temperature dynamics of the room, confirming the intuition



(a) Random PWM Input



(b) Heating Due to Occupancy



(c) Experimental (Solid) and Simulated (Dashed) Temperature

Figure 5.11: At each discrete time step, we applied a randomly generated input, which is the duty cycle of the PWM over 15 minute periods, taken from a uniform distribution ranging over $[0, 0.5]$. This was done over a period of the day (12:00AM to 5:30AM) during which the room is both in and not in use. Using semi-parametric regression [106], we identify both a discrete time model (5.4) and the term $q[n]$ which is given in units of $(^{\circ}C)$ and incorporates heating due to occupancy, equipment, and other external inputs. The measured room temperature is given in units of $(^{\circ}C)$ by the solid line, and a simulation of our model in units of $(^{\circ}C)$ is shown by the dashed line. The simulation uses the same inputs as provided to the BRITE platform over this range; the initial condition of the simulation is taken to be the experimentally measured temperature. The simulation has a root-mean-squared (RMS) error of 0.10 $^{\circ}C$.

of Ma, *et al.* [87] and results of previous work [11]. The function $q[n]$ has an average value of $6.98^{\circ}C$, and it is highly nonlinear with respect to time: It varies by up to $0.61^{\circ}C$ depending on what time of day it is. Furthermore, there are fluctuations over both long and short time horizons.

The heat generated by occupancy and equipment $q[n]$ displays interesting features. The room is a computer laboratory used by students at their own convenience and shows characteristics consistent with this role. The heat input $q[n]$ increases from lunchtime and peaks at 1PM, while the outside temperature peaks at 2PM. The occupancy has quick changes in

its direction at 3PM and 5PM. Finally, it is relatively constant from 8PM to 5AM, which is typically when there are few or no students in the room.

The large fluctuations have a major impact on the design process of a control scheme. This is because the nominal model for which a controller is designed can be inaccurate by $0.61°C$ (in our case) due to varying levels of occupancy. This causes issues with respect to efficiency, because standard MPC requires accurate models to provide high performance. For this reason, we make use of learning-based MPC [8] to design the controller. This MPC variant can estimate occupancy by measuring the temperature of the room and comparing it to what is expected by the model (5.4).

5.3.1.3.4 Modeling the Two-Position Control of Thermostat For the purpose of comparing the energy consumption of different control strategies, we must identify a model of the two-position control of the thermostat. The thermostat does its control in continuous time, and so this model is an ordinary differential equation. Part of the model is derived from a statistical analysis of temperature data from BRITE gathered over a 20 hour period. On average, the thermostat turns the AC on when the temperature reaches $22.8°C$ (standard deviation of less than $0.1°C$), and it turns the AC off when the temperature reaches $22.4°C$ (standard deviation of $0.1°C$). The thermostat has a feature called a heat anticipator that adjusts the top and bottom temperature thresholds, in an effort to conserve energy and reduce overcooling. We do not model this behavior. Furthermore, it takes the AC an average of 354 seconds (standard deviation of 75 seconds) to stop cooling the room after it is turned off. Though this is due to the internal dynamics of the heat pump, we approximate this by assuming that the AC stops cooling after a fixed time.

We again used semi-parametric regression on data from BRITE to estimate a continuous time model for the two-position control. The time constants for the room and heat transfer to the outside were taken from the discrete time model (5.4) and converted into continuous time constants by doing the reverse of an exact discretization. The model identified is

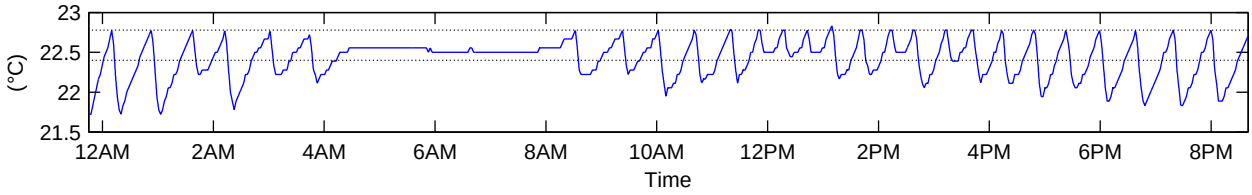
$$\dot{T} = -5.0 \times 10^{-4} \cdot T + 1.4 \times 10^{-4} \cdot w(t) - 1.2 \times 10^{-3} + q(t), \quad (5.5)$$

if the AC is turned on or for the first 354 seconds after it is turned off. Otherwise, the dynamics are given by

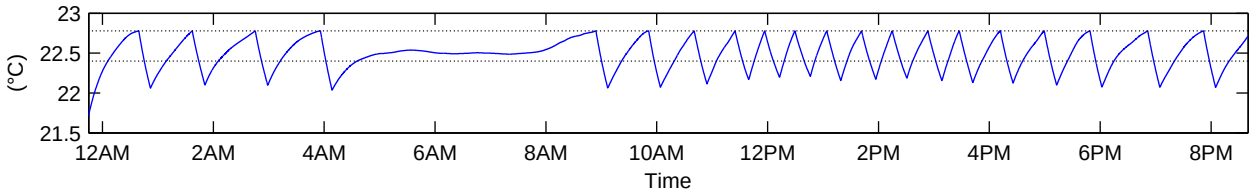
$$\dot{T} = -5.0 \times 10^{-4} \cdot T + 1.4 \times 10^{-4} \cdot w(t) + q(t). \quad (5.6)$$

In our model, the AC turns on when the temperature exceeds $T_{on} = 22.8°C$, and it turns off when the temperatures goes below $T_{off} = 22.4°$.

Visually examining the measured (Figure 5.12(a)) and simulated (Figure 5.12(b)) temperatures under two-position control indicates that there are several modeling errors; many of these are previously mentioned, but we collect them into one location for clarity. The temperature in the simulation rises slower than on BRITE, indicating that the identified



(a) Experimental Temperature



(b) Simulated Temperature

Figure 5.12: The thermostat used two-position control to maintain the temperature. Its average on and off temperatures were $T_{on} = 22.8°C$ and $T_{off} = 22.4°C$; these are shown by the solid, horizontal lines. The experimentally measured temperature is shown in ($°C$). Semi-parametric regression was used to identify a continuous time model. A simulation of this model using experimentally measured temperature as the initial condition is shown in ($°C$). The control of the simulation, which differs from the experimental control, was determined using the T_{on} and T_{off} values. The energy estimated by the simulation is $9.0kWh$, and (5.1) applied to the measured inputs computes $8.6kWh$; this is in contrast to the measured consumption of $8.6kWh$. Despite modeling errors, the energy estimates differ from the true value by only 5% and less than 1%.

time constant is slower than it should be. Furthermore, the model does not incorporate the internal dynamics of the heat pump or the thermostat's heat anticipator logic. Also, there is variation in the steady-state and transient energy consumption that is not captured in (5.1), which is used to make energy estimates.

Despite the modeling errors and simplifications, the simulation and (5.1) are reasonable proxies. The measured (Figure 5.12(a)) and simulated (Figure 5.12(b)) temperatures of the BRITE platform under two-position control over a period starting at midnight share similar qualitative features. The occupancy heating $q(t)$ is not shown because it displays characteristics similar to Figure 5.11(b) and that in previous work [11]. Moreover, the energy consumed by the AC was measured to be $8.6kWh$, computed by (5.1) to be $8.6kWh$, and simulated to be $9.0kWh$. This represents an error of less than 1% and 5%, respectively.

The overshoot of going below T_{bot} seen in both the measured and simulated temperature is in some sense wasted energy because it represents overcooling of the BRITE space. And even though the thermostat in BRITE has a heat anticipator that adjusts the T_{on} and T_{off} of the two-position control, it cannot adequately compensate for variations in weather and

occupancy. The learning-based MPC scheme we have developed can compensate for these factors, and so it can prevent overcooling and thus save energy.

5.3.1.4 Learning-Based Model Predictive Control of BRITE

Safety and robustness can be guaranteed with approximate models, but maximum efficiency requires accurate models. This tradeoff has driven research in adaptive control [6, 108] and learning-based control [2, 119, 128]. Statistical methods by themselves cannot ensure robustness [7, 10], and so the approach of learning-based MPC [8] is to begin with an approximate model of the system and refine it with statistical methods. Learning-based MPC is a rigorous control method that can (a) handle state and input constraints, (b) optimize system performance with respect to a cost function, (c) use statistical identification tools to learn model uncertainties, and (d) provably converge.

The control situation is as follows. We have a model (5.4) for the cooling dynamics of the BRITE room, and we have constraints on the maximum ($24°C$) and minimum temperature ($20°C$) to ensure comfort for people in the room. Preliminary experiments [11] made use of tube-MPC [34, 78, 81] (a form of robust MPC [16]) to ensure that these constraints were never violated despite varying occupancy and uncertainties in the weather forecast. However, testing over an extended period of time showed that the robust MPC described in [11] was too conservative because, when tracking a desired temperature of ($22°C$), the temperature rarely approached the constraints.

Consequently, we began to test standard linear MPC for its ability to stay within the desired temperature range. This control scheme had the same property in our tests—that it kept the room temperature within the constraints. It is important to remember this fact that a standard linear MPC ensures constraint satisfaction. However, the energy efficiency of this base scheme was lacking. It was unable to stay close to the desired ($22°C$), and it could use more energy than two-position control of the thermostat. Because of this, we implemented a learning-based MPC technique to control the room temperature.

5.3.1.4.1 Special Case of Learning-Based MPC The main idea of this technique [8] is that we decouple performance from robustness. By robustness, we mean whether an MPC scheme can ensure constraint satisfaction despite modeling errors and other uncertainties. Linear MPC itself has certain robustness properties [57]. As a practical issue, our tests on the BRITE testbed show that standard linear MPC gives sufficient robustness. In more general cases, we would need to use tube-MPC to ensure enough robustness for learning-based MPC [8].

We use a tilde to denote the temperature predicted by the learning-based model, an overline denotes the temperature predicted by the linear model with constant occupancy term $6.98°C$, and no overline indicates the measured temperature. The control action at time m, with temperature $T[m]$, control horizon $N = 20$ (5 hours), weight $p = 0.075$, and desired temperature $T_d = 22°C$ is given by the minimizer to the following optimization

problem

$$\begin{aligned}
\min_{u[\cdot]} \ & \textstyle\sum_{k=0}^{N} p \cdot (\tilde{T}[m+k] - T_d)^2 + && (5.7)\\
& + \textstyle\sum_{k=0}^{N=1} (r+\lambda) \cdot u[m+k] \\
\text{s.t. } & \tilde{T}[m+i] = 0.64 \cdot \tilde{T}[m+i-1] - 2.64 \cdot u[m+i-1] && (5.8)\\
& \qquad + 0.10 \cdot w[m+i-1] + \hat{q}[m+i-1] \\
& \overline{T}[m+i] = 0.64 \cdot \overline{T}[m+i-1] - 2.64 \cdot u[m+i-1] \\
& \qquad + 0.10 \cdot w[m+i-1] + 6.98 && (5.9)\\
& \overline{T}[m+i] \in [20, 24] && (5.10)\\
& u[m+i-1] \in [0, 0.5] && (5.11)
\end{aligned}$$

for $i = 1, \ldots, N$. The problem (5.7) generates an input $u[m]$ that minimizes the expected future performance of BRITE with respect to a cost function that encodes energy consumption and temperature deviation. Here, the term $\hat{q}[n]$ represents the predicted amounts of occupancy and computed using learning. In our unoptimized MATLAB code, this computation (5.7) takes between one to two seconds. It is easily scalable to larger problems because (5.7) is simply a quadratic program.

The optimization problem (5.7) decouples performance and robustness in the following manner. Robustness is due to the constraints (5.9) which are nothing more than the identified model (5.4) with constant occupancy. Performance is due to the use of (5.8) in the cost function. The intuition is that the cost depends on the learned occupancy $\hat{q}$ through (5.8), and the control is chosen such that the MPC without learning that is sufficiently robust (5.4) would satisfy the temperature and control constraints.

There are several important things to note about this formulation (5.7). The cost function contains (a) $p \cdot (T[m+k] - T_d)^2$ that represents deviation from the desired temperature and (b) the convex energy (5.2). This explicitly controls the tradeoff between keeping the temperature close to a comfortable value and the amount of energy used, and the value $p = 0.075$ was chosen because it gives a good tradeoff. Also, the convex energy (5.2) encourages a tradeoff between minimizing switching and duration of keeping the AC on, as discussed in Section 5.3.1.2.2.

5.3.1.4.2 Learning Occupancy Estimating occupancy is a detailed process that requires combining models of human behavior with sensors [80]. The BRITE platform faces an additional challenge because the occupancy varies immensely over the span of days and weeks, depending upon when assignments and projects are due. Some of the occupancy, such as for assignments, will likely be periodic in nature; other occupancy, like for projects, is more irregular and harder to predict. Furthermore, we need to know the heat generated by occupants and their use of computer equipment in the room for the purposes of energy efficient control. The correlation between the number of individuals in the room and the heat load will likely vary depending upon how many computers are in use.

	Method	Switches	On Duration	Energy Measured	Energy Estimated
Two-Position Control Experiment	LBMPC	94	6.0 hrs		$23.6kWh$
	Linear MPC	96	7.9 hrs		$30.5kWh$
	Two-Position	71	9.2 hrs	$32.6kWh$	$35.1kWh$
LBMPC Experiment	LBMPC	81	3.3 hrs	$11.8kWh$	$13.3kWh$
	Linear MPC	70	2.0 hrs		$8.6kWh$
	Two-Position	38	9.2 hrs		$34.5kWh$

Table 5.2: Summary of Experimental and Simulated Energy Comparisons on BRITE

Instead of relating the number of individuals in the BRITE room to the heat load $q[n]$, we focus our efforts on estimating this $q[n]$ directly from the temperature measurements and our model (5.4). We use the estimate

$$\hat{q}[m+i] \;=\; T[m] - \Big(0.64 \cdot T[m-1] - 2.64 \cdot u[m-1] + 0.10 \cdot w[m-1]\Big), \quad (5.12)$$

for $i = 0, \ldots, N-1$. The intuition is that the occupancy heating $q[n]$ is the discrepancy in what the linear model without the occupancy term 6.98 predicts the temperature at the next time step is and what the actual temperature is.

The approach we take in this paper is to use the simplest possible estimate — more accurate estimates of $q[n]$ taking into account specific models will only improve the energy efficiency of the BRITE testbed. An obvious extension is to fit our estimates to curves of best fit (e.g., a line or parabola) to compensate for the time-varying nature of the occupancy. Other extensions are to incorporate models of human behavior and other sensors.

We used this estimate of the occupancy for several reasons. As mentioned, this is the easiest estimate in terms of modeling: We do not need to worry about how to model long and short term human behavior. Secondly, it is well-behaved. Extrapolations using curves of best fit can significantly over- and under-estimate on long time-horizons. Finally, this estimate is easy to compute and shows that the learning can be done in a scalable manner.

5.3.2 Evaluating the Energy Efficiency of Learning-Based MPC with BRITE

The original aim of building the BRITE platform [11] was to enable evaluation of existing methods and design new control schemes that minimize the energy consumption needed to maintain a comfortable temperature in the room. In our experiments, linear MPC had inconsistent performance due to its inability to compensate for the impacts of occupancy – at times it had difficulty with either saving energy as compared to two-position control [11] or maintaining temperature close to the desired value. This is related to a fundamental tradeoff in control systems between robustness to model uncertainty and performance due to model accuracy [8].

	Method	Tracking Error	Temperature Variation	Average External Load
Two-Position Control Experiment	LBMPC	$0.75°C$	$0.13°C$	$11.0°C$
	Linear MPC	$0.62°C$	$0.30°C$	$11.0°C$
	Two-Position	$0.61°C$	$0.20°C$	$11.0°C$
LBMPC Experiment	LBMPC	$0.86°C$	$0.17°C$	$7.2°C$
	Linear MPC	$0.93°C$	$0.21°C$	$7.2°C$
	Two-Position	$0.55°C$	$0.19°C$	$7.2°C$

Table 5.3: Summary of Experimental and Simulated Temperature Comparisons on BRITE

We implemented a learning-based MPC scheme [8] with the intuition that this could provide improved performance and reduce energy usage. Our experiments on the BRITE platform suggest that this is indeed the case. The energy improvements come from two features of the learning-based MPC. First, it can compensate for fluctuations in weather and occupancy through learning. For instance, the two-position control of the thermostat overcools the room when occupancy is low, and the heat anticipator in the thermostat does not adequately compensate. Second, it considers the penalty due to electrical consumption by the heat pump transient and tries to optimize the tradeoff between minimizing switching and AC on time.

5.3.2.1 Experimental Methodology on BRITE

A key challenge of an *in situ* platform, as opposed to an undisturbed testbed or simulator, is the inability to perform direct comparisons between control methods with identical system conditions. One potential solution, consecutive experiments, is insufficient due to the huge variability in weather and occupancy on the time scales of hours, days, and weeks.

To address this, we run one control scheme on BRITE, and then simulate others. The benefit of this is that identical weather and occupancy conditions allow for an equal comparison between the methods. The disadvantage is that there will obviously be some error between the simulated and real energy consumptions because of modeling mismatch. To mitigate this, we alternate which method is simulated between the new and existing control schemes, making the following comparisons:

- Run the two-position thermostat control on the BRITE testbed and simulate the linear and learning-based MPC using measured weather and occupancy.
- Run the learning-based MPC on the BRITE testbed and simulate the linear MPC and two-position thermostat control using measured weather and occupancy.

After simulating a control scheme, the corresponding energy consumption can be estimated by summing the steady state and transient electrical energy components (5.1). We present one instance of each comparison.

The results of two experiments are summarized in Tables 5.2 and 5.3. The first controlled BRITE with two-position control, and the second involved implementing the learning-based MPC on BRITE; we do not include any experiments with linear MPC on BRITE. The energy usage measured by BRITE and estimated by (5.1) are both provided for these experiments. Their performance is compared to the other control schemes though simulations that use identical weather and occupancy levels. The table lists the number of times the AC was turned on, the duty cycle of the AC, the tracking error as measured by the RMS error between the room temperature and the desired temperature $T_d = 22°C$, and the variation in the room temperature as measured by its standard deviation. The external load corresponds to the average temperature increase over 15 minutes caused by the weather and occupancy.

5.3.2.2 Two-Position Control Experiment on BRITE

Over a 24 hour span beginning and ending on a weekday, we started running the two-position control of the thermostat on BRITE at 11PM. The experimentally measured temperature is shown in Figure 5.13(a). Using our models, we simulated the corresponding behavior of the learning-based MPC, which is shown in Figure 5.13(b). For our simulation, we used the stored weather forecasts, true weather temperature, and occupancy estimated using our model of two-position thermostat control. The learning-based MPC used an estimated 28% less energy than the two-position control. The PWM control actions corresponding to two-position control and learning-based MPC are shown in Figure 5.13(c) and Figure 5.13(d), respectively. Moreover, Figure 5.13(e) shows the change in temperature over 15 minutes corresponding to experimentally measured weather and occupancy (i.e., $k_w w[n] + q[n]$).

5.3.2.2.1 Learning-Based MPC Experiment on BRITE We ran the learning-based MPC control on the BRITE platform over a time range that covered two weekdays, and started at roughly 1PM. The experimentally measured temperature is shown in Figure 5.14(a). Using our models, we simulated the corresponding behavior of the two-position control, which is shown in Figure 5.14(b). For our simulation, we used the true weather temperature and occupancy estimated using our model of the learning-based MPC. Our learning-based MPC approach on BRITE is estimated to reduce the energy consumption by 66%, when compared to the existing two-position control scheme. The control actions corresponding to the two-position and learning-based MPC are shown in Figure 5.14(c) and Figure 5.14(d), respectively. The measured weather and occupancy for this experiment $k_w w[n] + q[n]$ is given in Figure 5.14(e).

5.3.2.2.2 Discussion of Results Both comparisons show that significant energy is saved by the learning-based MPC scheme. It is useful to discuss what features of our implementation and scheme contribute to this, because many of these principles may generalize to other HVAC systems. Broadly speaking, the improvements come about through the use of modeling and statistical techniques.

Considering the discrete time version of a mathematical model taken from physical intuition (5.3) and identifying its parameters help improve the efficiency. As stated earlier, there are complex dynamics in the heat pump, and the evaporator continues to cool air for some time after the heat pump is turned off [124]. The discrete time form of the model (5.3) accounts for this behavior by considering the AC behavior over a 15 minute span of time, rather than the instantaneous behavior.

Furthermore, identifying the parameters of the model allows us to be able to estimate occupancy through only temperature measurements, as in (5.12). These occupancy estimates are important because this feature of the system adds considerable variation in the temperature dynamics of the room. Whereas two-position control overcools the room when there is lower occupancy, learning-based MPC detects lower levels of occupancy and reduces the amount of cooling.

Lastly, the electrical energy characteristics of the heat pump are important to conserving energy. The transients of the heat pump effectively add a penalty, in terms of energy used, for switching too frequently. The learning-based MPC can make a tradeoff between how long the heat pump is turned on for and how often it switches, and it can dynamically adjust this tradeoff based on the estimated occupancy.

This tradeoff is actually very interesting, because it leads to counter-intuitive behaviors with the learning-based MPC. Examining the temperature of the learning-based MPC (i.e., Figs. 5.14(a) and 5.13(b)) shows that the bands within which the temperature is maintained actually vary over time. Generally speaking, when the outside temperature or occupancy are high, the learning-based MPC actually tightens the temperature bands. When the outside is cold or occupancy is low, the learning-based MPC widens the temperature bands.

These behaviors can be explained by thinking of the electrical behavior of the heat pump. When the temperature or occupancy is high, the AC needs to be turned on for a greater fraction of time. The steady state energy consumption is much higher than the transient energy consumption, and so the learning-based MPC does not penalize as much for frequently switching. In fact, it increases switching to prevent overcooling the room. In the opposite situation, the AC needs to do less total cooling. Here, the steady state energy consumption is smaller and so transient energy due to switching becomes important. The learning-based MPC tries to reduce switching in these cases and allows for larger temperature variations in the room.

5.4 Electric Loads with Local Energy Storage

The previous supply-following loads have employed thermal mass as an energy storage element, exercising the slack inherent in thermostatic control. Supply-following loads can also be made from battery-based systems, where an additional energy storage element has been added to the system. In this section, we create supply-following loads from two battery-based systems: a refrigerator enhanced with an additional thermal storage element and a laptop that employs a Li-ion thermochemical battery.

5.4.1 A Refrigerator Augmented with Controllable Thermal Storage

In Section 4.2, we examined the effect of adding utility-scale energy storage to the electricity grid. An important implication of how the model of storage was designed is that energy storage is managed centrally, generally by the utility, and exists in the grid closer to the sources of generation. In this section, we examine a distributed model of energy storage, where the storage resides at grid endpoints, closer to locations of demand. In particular, we evaluate the class of flexible loads that incorporate energy storage, an underused and potentially highly valuable resource for the electric grid. Though they do not attract the vast attention and broad research community of electric vehicles or buildings, they have notable advantages over both. These types of loads, primarily refrigeration and freezer systems, are already widely deployed – over 99% of U.S. homes have a refrigerator, and turnover is relatively fast, with more than half of refrigerators being replaced within seven years of purchase, often with incrementally better technology. In fact, mass production and improved energy efficiency have resulted in a 6-fold improvement in power consumption over thirty years while lowering costs and growing capacity [33]). The relatively small number of layouts and assemblies presents a distinct advantage over buildings, which are nearly universally custom designed. Also, unlike battery storage, thermal storage has high turn-around efficiency and nearly infinite charge cycles. Further, in many cases, there are fewer conversions, as electricity that becomes thermal energy will not need to be converted back to electricity in order to be utilized. Aside from all its energy benefits, substantial thermal storage makes for a better, more robust fridge.

To evaluate the potential of refrigerators to provide grid-scale energy storage, we developed a prototype system that couples these relatively simple mechanical systems with a wireless sensor network, a thermal reservoir to enhance energy storage, and a modified control loop. For our study, we employ a typical domestic refrigerator, a Whirlpool/KitchenAid (Model #KTRD18KDWH00) unit, with a 13.1 ft^3 refrigerator compartment and a 5.0 ft^3 freezer compartment. Our sensor network consists of Type-K thermocouples in each of the refrigerator and freezer compartments as well as outside the refrigerator to measure ambient temperature. These sensors are connected to a LabJack UE9 [77] data acquisition device and are polled every 10 seconds. For measuring and external control of power consumption, we use an ACme wireless plug meter and load switch connected via a 6LoWPAN-compliant IPv6 edge router [69] reporting data every 5 seconds.

Though the volume of data involved in this monitoring is not so large as to be intractable, the diversity of data sources made reliable system monitoring somewhat challenging. One tool we made heavy use of was the Simple Measurement and Actuation Profile (sMAP) system [40]. sMAP provides a JSON web services interface to a highly engineered time-series database system, essential for maintaining, archiving, and organizing the array of data streams we collected, both locally from the sensors we deployed as well as from remote sources like weather data, grid energy generation, and real-time electricity pricing from the Internet.

From Section 5.2, which examines the slack in the energy load of refrigerators, we know that the thermal mass of the contents of a refrigerator can be used to maintain an acceptable temperature range in both the refrigerator and freezer compartments without running the compressor. Though this load flexibility is a useful and largely untapped energy storage resource, its magnitude is relatively small per instance, shifting low hundreds of watts by low tens of minutes. This capability comes at the cost of installing the extra communications and sensing capability needed to exploit it, likely not economically feasible even over a large population of refrigerators. In order to improve the utility of this flexible resource at grid-scale, more energy storage per unit and more precise temperature control is needed. This allows for more slack in the energy load, and thus more capability to either avoid electricity consumption at undesirable times or consume electricity at surplus occasions, perhaps due to excess local or grid renewables generation.

To address this, we constructed a prototype thermal storage system, consisting of three sealed tanks placed into the freezer compartment and filled with a phase change material (PCM) for energy storage. A picture of the setup inside the freezer compartment is provided in Figure 5.15. The PCM used is the eutectic concentration of aqueous ammonium chloride, which has 19.7% NH_4Cl by mass, is nontoxic, and has a eutectic point of -15.4 ° C, just above the average operating range of our freezer. The frozen PCM can be opportunistically melted to avoid operating the compressor and consuming electricity. When electricity is available again, the partially- or fully-melted PCM can be re-frozen and charged for its next use. Use of a PCM allows far more energy to be stored via the latent heat rather than only the sensible heat from a change in temperature.

To aid in heat transfer, one of these tanks has non-corroding stainless steel tubing coiled inside which is in turn connected to a heat exchanger in the refrigerator compartment, and a small fluid transfer pump that is outside of the refrigerator, creating a loop that uses a nontoxic propylene glycol-based solution with a very low freezing point to transfer heat. To further encourage heat transfer, the storage tanks have a high surface area-to-volume ratio and two 1 W fans are placed in the refrigerator compartment to circulate air over the fins of the heat exchanger.

The three tanks in our setup have a total of 10 kg of NH_4Cl salt in an aqueous solution. To calculate how much energy can be stored in this medium, we consider the *enthalpy of fusion* of our salt compound, a property describing the magnitude of heat transfer during the solid-liquid phase change. For NH_4Cl, this quantity is 10.6 kJ/mol [36], meaning that 10 kg of salt solution is able to store 0.55 kWh of energy. Given the power consumption of the refrigerator, this energy store should maintain fridge temperature for 6.35 hours. Section 5.4.1.1 provides further empirical evidence that corroborates this calculation.

A comparison of the performance of the refrigerator with and without the thermal storage element is provided in Table 5.4. A trace of operation of the refrigerator with the thermal storage device is in Figure 5.16. From the picture of our prototype, it is obvious that this is an early design; though we attempted to at least maintain a semblance of the structure of the freezer compartment, we fully believe that a commodity design of this type of system would be highly integrated into the refrigerator, incorporated with its internal control system and

possibly providing insulation in addition to thermal storage capability.

	Unmodified Fridge	**Fridge with Energy Storage**
Avg. Power	76.69 W	86.55 W
Duty Cycle	53.32%	60.14%
Mean Compressor Duration	1558 s.	1762 s.
Mean Defrost Duration	11.4 mins.	11.4 mins.
Mean Heating Post Defrost	13.5 mins.	3.1 mins.
Defrost Frequency	15 hrs.	13.7 hrs.

Table 5.4: Comparison of power consumption and cycle behavior of two fridge configurations. Values are calculated over a 12-day, unperturbed period.

5.4.1.1 Applications Using Enhanced Thermal Storage

Though the increased thermal storage of this prototype is not novel, pairing it with a sensor network and modified control loop enables new and potentially valuable applications. In this section, we focus on two key applications, price-responsive demand and supply-following, that explore questions about communications requirements and the benefits of this technology to customers and utilities. We implement each application using our prototype, and then employ models of the California grid at varying levels of renewables penetration to evaluate the benefits of this technology, both now and in future grids.

5.4.1.1.1 Price-Responsive Demand To combat the enormous costs of peak electricity and provide flexible consumers with potential benefits, electricity utilities are increasing the availability of time-of-use (TOU) electricity tariffs. These tariffs set electricity prices based on a fixed schedule, meaning that communication with the grid or the Internet is not necessary – the system only needs to know the time and the schedule. Table 5.5 represents two different summer (May-October) weekday TOU pricing schedules offered by the California utility Pacific Gas & Electric (PG&E), the A6 schedule for small commercial customers (with a monthly average of 200-500 kW) and the E6 schedule for residential customers [98]. Each schedule has a peak pricing period that lasts for six hours (12-6 PM for A6, 1-7 PM for E6). By operating the thermal storage device to shift consumption from these hours, as shown in Figure 5.17, consumers can save on electricity bills.

The amount of savings depends substantially on the pricing tariff, which may change by season. In the case of PG&E, with our thermal-storage refrigerator, a residential customer can save \$0.04 per summer weekday, amounting to a nearly \$5 reduction in an overall summer weekday electricity bill – a 13% reduction versus an unmodified refrigerator. However, the extra electricity consumption needed throughout the year to maintain the thermal storage

TOU Tariff	Summer Weekday Hours	$/kWh
Residential E6		
Off Peak	9:30PM-8:30AM	$0.137
Partial Peak	8:30AM-12PM, 6PM-9:30PM	$0.225
Peak	12PM-6PM	$0.444
Commercial A6		
Off Peak	9PM-10AM	$0.098
Partial Peak	10AM-1PM, 7PM-9PM	$0.170
Peak	1PM-7PM	$0.279

Table 5.5: Pacific Gas & Electric commercial and residential summer weekday time-of-use electricity tariffs.

counteracts this benefit, resulting in zero net annual savings. The same holds true for commercial customers, except for one notable difference. Commercial TOU customers often also pay a demand charge, reflecting their total peak demand over the period. For PG&E, the fee is over $13 per kW. Thus, the potential that a defrost cycle, consuming 0.4 kW, may occur while other electricity consumption is at peak may incur an additional demand charge of up to $5.20 for a single month, and possibly over $60 in a year (or more with a larger refrigerator). Though these benefits to consumers are modest, as we explain in Section 6.1.2, utilities can derive significant benefit from price-sensitive consumers.

5.4.1.1.2 Supply-Following As discussed in Chapter 3, the emergence of large amounts of renewable generation is a critical shift for the electricity grid. Flexible energy loads like the thermal storage system we demonstrate can act as supply-following loads, advancing or delaying consumption to better match fluctuating electricity supplies, as introduced in Section 4.3.1. This supply-following capacity can play a huge role in accommodating the increased variability presented by emerging renewable sources.

Towards that end, we present a prototype implementation of a supply-following refrigerator. As opposed to a traditional refrigerator consuming energy irrespective of conditions external to its physical control loop, a supply-following refrigerator is able to use its communicative ability to monitor grid conditions, avoiding electricity consumption when it is least available. Though it is possible to use more advanced forms of control and consider a wider array of factors, we begin with a simple formulation, giving our supply-following refrigerator three operating modes: low supply, medium supply, and high supply. The fridge selects its mode of operation by observing a real-time feed of total renewables generation from the California ISO, composed primarily of solar and wind. Based on the supply level, it can choose to delay high-power defrost cycles and maintain a slightly warmer temperature in low supply mode, operate normally in medium supply mode, or run with shorter, more frequent, and higher energy cycles in high supply mode. Figure 5.18 shows the behavior of the fridge and its thermal storage system over three days, providing the average power consumption experienced in each mode.

The results show how dynamic a supply-following refrigerator can be as compared to

a traditional refrigerator, with the low supply mode consuming less than 60% the average power of the high supply mode. Further, the fridge operates on precisely the opposite pattern that today's grid operators favor – it consumes little energy at night when the sun is not shining, but it wakes up and consumes more during the day when there is an increase in solar energy production. The fridge is configured to select its mode of operation based on thresholds of the aggregate renewables signal, with low supply mode below 3500 MW, high supply mode above 4500 MW, and medium supply mode in between; this selection provides a range of operating modes over the course of three days, but selection of operating modes could certainly be enhanced with forecasting and response to grid demand. We believe this is the first instantiation of a real-time supply-following refrigerator, responding directly to current energy availability.

5.4.1.2 Deployment Considerations

There are many practical considerations in the eventual deployment of this technology. First, we emphasize that the design in this study is a prototype implementation. A commodity-scale implementation would likely have a far smaller physical footprint, perhaps integrating into the structure of the refrigerator itself, would use lower power pumps and fans or discard them entirely, and would have improved control over refrigerator processes like defrost cycles. Additionally, it is also important to consider occupant interaction with refrigerators and the effect on load; since door opening and closing events can occur randomly but tend to be correlated with mealtimes, this would have an effect on refrigerator load [70]. However, opening a refrigerator door has a far more substantial effect on air temperature than it does on the temperature of contents, which have far higher thermal mass; we expect that the thermal storage device would see very little change from increased door opening activity. Another important concern is adoption; utilities gain more benefit from this type of product, but it is consumers that need to purchase and replace these products. In the past, utility and federal rebates and licensing programs (*e.g.*, EnergyStar) have helped to accelerate this process. A final concern is synchronization across devices throughout the grid. Though it is possible that many devices could turn on and off simultaneously, this problem would easily be solved by adding jitter into the controller's operation.

5.4.2 Supply-Following Laptops

Another viable area for creating supply-following loads is systems with thermochemical energy storage. In this section, we create a charging controller for a set of laptop computers, which each incorporate a Li-ion battery. [2] Though the overall consumption of this class of load is small, our techniques abstract to much larger batteries, such as those found in grid backup systems or electric vehicles (EVs) [17]. At this point in time, though penetration of grid-scale batteries and EVs remains modest, their prevalence is quickly increasing; though

[2]The study on laptops was carried out in collaboration with Nathan Murthy.

global sales of battery-powered vehicles are roughly 2 million during the year 2013, this number is expected to grow to 7 million annually by 2020 [94]. Further, these batteries are relatively much larger than laptop batteries; for example, EV battery picks are composed of hundreds or thousands of Li-ion cells. Many of the same battery-oriented observations and charge/discharge management techniques used in adapting an array of laptops to be supply-following are relevant for grid-scale batteries and EV batteries as well. In this study, we used laptops as a proxy for these larger batteries, as they are much easier to obtain, characterize, and manage.

5.4.2.1 Laptop Charging Model

As in other cases, in order to create supply-following laptops, we must derive models for operation of the constituent parts of the laptop, particularly the battery. Our models fundamentally rely on understanding the power consumption profiles of laptops as a function of their capacities or states of charge (SoC). It is already well known that as the charge in a lithium-ion battery is depleted, the resistance within the battery increases and therefore the amount of power delivered to a battery in a low SoC is higher than the amount of power delivered to a battery in a high SoC. It is also well known that the charging dynamics and capacity limits of rechargeable batteries vary over their lifetimes depending on the nature of their charge and discharge cycles. Without even introducing vendor-specific laptop designs, these simple facts about laptop batteries alone imply that our problem cuts across a great deal of heterogeneity which makes this problem particularly interesting from a characterization standpoint.

In general, we can classify laptops as either having or lacking charge control mechanisms. These built-in charge controllers, or lack thereof, affect the individual power consumption patterns aggregated over each subcomponent of the said laptop. The power traces in Figure 5.19 reveal this classification empirically. The graph on the left (Figure 5.19(a)) depicts power and capacity traces over one full charging cycle of a laptop possessing a charge controller that produces a decreasing stair-step power trace. In Figure 5.19(b), we depict a charging cycle of a laptop that lacks such a controller, and so we observe an exponentially decaying power trace profile typical of lithium-ion batteries. In this work, we employ a model that does not employ a charge controller; the continuous monotonically decreasing power required to charge a battery lends itself well to optimization. It would also be possible to model a charge control mechanism by employing a state-based model.

Modeling the total power consumption of laptops with all their various interconnected subcomponents is more complex than modeling only a battery. The dynamic nature of operating systems and software means that the electronic subcomponents on which they run also behave dynamically (as demonstrated by the seemingly stochastic high-frequency variations of the above power traces), and thus the complexity of managing the total power consumption of a large number of these devices is compounded as we include a load management policy specific to each laptop for each subcomponent in addition to each battery.

In order to understand how each measurable subcomponent contributes to the electrical power consumption of the laptop as a whole, we use the same laptop that produced the results in Figure 5.19(a) to estimate which subcomponents contribute to the bulk of this power consumption. Although some software tools [52, 130] provide insights into how application-level activity affects power consumption of software programs, we use empirical methods that reveal how low-level hardware activity, as opposed to application-level activity, affects power consumption. We disaggregate the peak power consumption of the subcomponents in Figure 5.20 using the same measurement techniques employed to isolate the power delivered to a MacBook lithium-ion battery that is charging. For components such as the network interface card (NIC), LCD screen, and built-in cooling fan, we throttle parameters – NIC bandwidth caps, screen brightness, and fan speed, respectively – to get a clear sense of how much power is delivered to these components from their minimum to maximum operation thresholds. The subcomponent disaggregation figure shows that power delivered to this particular laptop battery in low-capacity states accounts for up to 56.1% of total laptop power consumption during charging.

Given the above considerations and in order to reduce the complexity of the problem, we tailor our algorithms around developing load management policies that manage the charging schedules of only laptop batteries since they account for the bulk of power delivered to laptops when drawing power from an outlet. More importantly, knowing the functional relationship revealed in Figure 5.19 between the power consumption of the laptop and its battery capacity state allows us to project in advance how much power a laptop is likely to consume given the SoC of its battery even if it is not presently drawing power from an outlet.

5.4.2.2 Problem Formulation

We investigate two model-predictive control problems and develop algorithms that attempt to solve each of these distinct but related problems. In each of the problems, we simulate the scenario of laptops entering and leaving a building. The first control problem, which we call the *classic demand response problem*, involves reducing absolute power consumption from a baseline during a demand response event. A demand response event (DR event) is a period of time during which the available supply is below the reserve margin set forth by a balancing authority, and so the total load on the grid could exceed supply if action is not taken to curtail demand. With respect to our first scenario, we develop a scheduling algorithm whose objective is to minimize the aggregate load of laptops randomly entering and leaving a building that has been issued a DR event in order to reduce the building's overall peak load during the event.

The second control problem, which we call the *continuous demand response problem*, involves fitting a random load with a random supply over a continuous rather than discrete period of time. Just as in the previous scenario, the random load is the power consumption of laptops entering and leaving a building. But rather than being issued DR events as in the first scenario, this building comes locally equipped with a variable energy resource (*e.g.*,

rooftop solar). The solution to this problem relies on techniques that fit load to supply in a manner that optimizes the use of these variable resources both in times of scarcity and excess by minimizing reliance on external power.

Irrespective of either control problem, we wish to have an accurate model of the load we are simulating in each demand response scenario. To achieve this, in addition to building an accurate battery model, we also construct a weighted Poisson arrival and departure process to simulate laptops entering and leaving our control area based on building occupancy inferences from empirical load profile data.

In both DR scenarios we begin with the initial set $\mathcal{S}_0$ of N laptops at t=0 on the domain $[0,\ T]$. For each laptop $\ell_1, .., \ell_i, ..., \ell_N \in \mathcal{S}_0$ we seed their capacity traces so they are randomly chosen to either have one or the other charge profile observed in Figure 5.19 to introduce some reasonable load heterogeneity in our model and parameterize the battery capacity X_i of the i^{th} laptop as

$$X_i(t+1) = X_i(t) + u_i\mathcal{C}_i(X_i(t))\delta t + (1-u_i)\mathcal{D}_i\delta t \tag{5.13}$$

where $u_i \in \{1, 0\}$ represents whether the i^{th} laptop battery is charging ($u_i = 1$) at a rate $\mathcal{C}_i(X_i(t))$ (*e.g.* the derivative of either of the capacity traces from Figure 5.19) or discharging ($u_i = 0$) at a rate $\mathcal{D}_i$ as it transitions from t to $t+1$ with a step size of δt. Unlike the charge rate which is a function of the capacity of the battery, we observe (also from empirical measurements) that the discharge rate of laptop batteries is mostly constant in both high and idle activity states. Furthermore, each laptop in $\mathcal{S}_0$ has the initial power traces

$$u_1P_1(X_1), ..., u_iP_i(X_i), ..., u_NP_N(X_N).$$

In our models, we initialize the capacity states, and hence the power states, of each laptop so that they are uniformly random on $[0, X_{i,max}]$.

We then introduce a weighted Poisson arrival and departure process to simulate laptops entering and leaving the control area (or zone) with arrival and departure rates defined as

$$\begin{aligned} \lambda_{arrival} &= w_t\lambda \\ \lambda_{departure} &= (1-w_t)\lambda \end{aligned} \tag{5.14}$$

for a fixed λ whereby $w_t \sim \mathcal{L}_{building}$ and where $\mathcal{L}_{building}$ is the inferred occupancy of the building based on its load profile which roughly provides an estimate of laptops in the zone. In this way, as building occupancy increases, the rate of laptops arriving increases, and the rate of laptops leaving decreases. And conversely, as the occupancy decreases, the rate of laptops arriving decreases, and the rate of laptops leaving increases. So then the set of laptops in subsequent time steps increases or decreases according to

$$\mathcal{S}_{t+1} = \mathcal{S}_t + arrivals_{t+1} - departures_{t+1}, \tag{5.15}$$

Thus over each time step on the domain, the total number of laptops in each set $\mathcal{S}_t$ varies across the chain of step transitions

$$\mathcal{S}_0 \to \cdots \to \mathcal{S}_t \to \cdots \to \mathcal{S}_T$$

as laptops enter and leave the zone in this aleatoric fashion.

5.4.2.2.1 Classic DR Model In the classic DR problem, the goal is to reduce peak demand by some percentage of baseline power consumption during a scheduled DR event that lasts for some known duration. We call this percent reduction the *curtailment coefficient* of the DR event which occurs on the interval $[t+k, T-m]$ for $t+k < T-m$. In real world grid operations, the event typically lasts between 3-6 hours.

For this scenario, our model tries to find a combination of laptops $\ell_{1'}, ..., \ell_{k'} \in \mathcal{S}_t$ at each time step t that minimizes the aggregate peak power consumption over the entire event duration. We construct a *time-varying bounded knapsack* algorithm to accomplish this whereby for each $t \in [t+k, T-m]$ we try to find a subset $s_t \subseteq \mathcal{S}_t$ that optimizes the curtailment coefficient c_t thereby minimizing the peak power consumption in the interval.

Since laptop arrivals and departures are random, we can, at best, only try to predict the curtailed load during the demand response event. In real implementations, we would take historical power data of a building with densely deployed networks of laptops paired with outlet sensor-actuators and then perform machine learning operations on the data to extract the patterns needed to make baseline and curtailment load projections. For our purposes, it is sufficient to simulate arrival and departure rates described in Equation 5.14 to produce load projections. We make a curtailed load projection $\sum_{\ell_i \in s_t}^{s_t} u_i P_i$ on $[t+k, T-m]$ with $u_i \in \vec{\mathbf{u}}_t$ where $\vec{\mathbf{u}}_t$ is a vector whose elements correspond to whether the i^{th} laptop is charging or not. The elements of $\vec{\mathbf{u}}_t$ are assigned either 0 or 1 (in real deployments, entries of $\vec{\mathbf{u}}_t$ correspond to the actuation state of sensor/actuator nodes) by choosing the optimal $c_t \in (0, 1]$ as $t : t+k \rightarrow T-m$ by carrying out the following knapsack formulation:

$$\text{objective} : maximize\ c_t \in (0,1]$$
$$minimize\ max\left\{ \sum_{\ell_i \in s_t}^{s_t} u_i P_i \right\}$$

such that

$$\sum_{\ell_i \in s_t}^{s_t} u_i P_i \quad \leq \quad (1 - c_t) \sum_{\ell_i \in \mathcal{S}_t}^{\mathcal{S}_t} P_i \tag{5.16}$$

Unfortunately, the search space for assigning elements in $\vec{\mathbf{u}}_t$ with this algorithm is very large. To reduce this space, we introduce a method of prioritizing laptops weighted against their power and capacity states which we call a *z-score* defined as

$$\begin{aligned} z_i(X_i(t), P_i(t)) = \alpha & \left(1 - \frac{P_i(t)}{P_{i,max}} \right) \\ & + \beta exp\left(-\kappa \frac{X_i(t)}{X_{i,max}} \right), \end{aligned} \tag{5.17}$$

where α, β, and κ are chosen so that the X_i term dominates z_i for small values of $X_i(t)$ (*i.e.* low battery capacity states). In this equation, $P_{i,max}$ and $X_{i,max}$ are the largest observed

values of laptop ℓ_i. The z-score represents a function that increases as the SoC of a laptop battery diminishes or as the total power draw of the laptop decreases. The higher the z-score, the higher the laptop's charging priority in the scheduler. Since laptops in low capacity states will also likely draw more power than laptops in higher capacity states, we pick α and β so that $\alpha < \beta$ in order to weight capacities over power consumption levels (κ must also be positive to ensure that z is monotonically decreasing on $[0, 1]$). Therefore, the laptop in the lowest capacity state and lowest consumption level has the highest charging priority, and the laptop with the highest capacity and highest consumption level has the lowest priority.

To shrink the search space, we sort laptops in $\mathcal{S}_t$ according to z-scores in descending order. We then iterate over this ordered set and pick the ℓ_i's that satisfy the above constraint. We summarize the sort-and-assign procedure as follows:

1. Sort $(\ell_1, ..., \ell_N) \in \mathcal{S}_t$ from $z_{max} \rightarrow z_{min}$ which yields an ordered set $(\ell_{1'}, ..., \ell_{N'}) \in \mathcal{S'}_t$
2. Pick $(\ell_{1'}, ..., \ell_{k'}) = s_t \subseteq \mathcal{S'}_t$ that solves the time-varying bounded knapsack problem in (4) by assigning the corresponding elements of $\vec{\mathbf{u}}_t$ to their appropriate boolean values so that each laptop with $u_i = 1$ is allowed to charge (and each with $u_i = 0$ is disallowed from charging) in $\mathcal{S}_{t+1}$

Last, we may prefer a policy that does not allow any laptop battery to die. In this case, as in our model, we introduce a capacity acceptance margin $X_{i,accept}$ whereby any laptop ℓ_i that satisfies $X_i(t) \leq X_{i,accept}$ is allowed to charge. By introducing an acceptance threshold, we would not want to have laptops marginally break above $X_{i,accept}$ immediately after time δt of charging only to fall below $X_{i,accept}$ again after discharging for δt in the next time step. To mitigate this, we also introduce a rejection threshold $X_{i,reject}$ that will only disallow laptops from charging once they have reached $X_i(t) \geq X_{i,reject}$. This will prevent oscillations across capacity thresholds. Alternatively, we might have z-score acceptance and rejection thresholds derived from $X_{i,accept}$ and $X_{i,reject}$ so that we can use the same sort-and-assign procedure as above. We implement the latter policy in both the classic and continuous DR control problems in simulations we run.

We show an example of a simulation run of this model in Figure 5.21 that achieves a more than 60% average reduction of load from the baseline during a 3-hour DR event. We quantify the deferrability of the whole cluster of laptops as a single load during the event by tracking the *energy slack* [114] of the system under control in the bottom-right chart of the figure.

5.4.2.2.2 Continuous DR Model In the continuous DR scenario we try to fit a time-varying random load to a random renewable supply. Unlike the classic DR problem where the goal is to reduce peak power consumption over the duration of a DR event, the continuous scenario involves matching load with supply in times of excess and scarcity over the entire domain $[0, T]$ during which some intermittent resource is generating electricity albeit variably. The classic DR model assumes that all power delivered to each load originates from the grid. However, in the continuous model we assume that the intermittent supply is at the site of

consumption with the grid available as an external resource. The goal, then, is to minimize dependence on any external or supplemental power supply.

We construct the continuous model using the same techniques as the classic model with a nearly identical problem formulation. In this time-varying bounded knapsack algorithm, there exists only one objective function, namely, the absolute error of aggregate load and total supply over each time step $t \in [0, T]$ under the constraint that load does not exceed supply. The formal problem statement is

$$\text{objective}: \ minimize \Big| \sum_{\ell_i \in s_t}^{s_t} u_i P_i - \hat{P}_{supply}(t) \Big|$$

$$\text{such that} \sum_{\ell_i \in s_t}^{s_t} u_i P_i \ \leq \ \hat{P}_{supply}(t) \tag{5.18}$$

where $u_i \in \vec{\mathbf{u}}_t$ and $\hat{P}_{supply}(t)$ is the predicted supply of power at time t. In this way, part of the performance of this supply-following algorithm depends on the accuracy of the predictor. We quantify the error between load and supply as the *grid dependence*, defined on a real-valued scale of [0,1] with 0 being no dependence on external power whatsoever and 1 being total dependence.

The continuous scenario employs the same sort-and-assign procedure to populate the entries of $\vec{\mathbf{u}}_t$ by using the z-score of each laptop to prioritize them for charge scheduling in each time step but with only a single objective function to fulfill as opposed to two objective functions in the first scenario. When no renewable power supply is available (*e.g.* night time in the case of solar), the z-score acceptance and rejection thresholds are put into effect to ensure that no laptop dies at the expense of grid independence.

5.4.2.3 Results from Laptop Simulation Studies

We use the same z-score function (with $\alpha = 0.25$, $\beta = 0.75$, and $\kappa = 3.0$) to prioritize laptop charging schedules for both the classic and continuous DR simulations. A key metric we test in the classic DR scenario is how well our curtailment algorithm performs over varying DR event durations since the available aggregate energy slack of our loads towards the end of every DR event approaches zero as the battery capacity of each laptop depletes (recall Figure 5.21). Less slack implies less deferrability and therefore less curtailment feasibility. We summarize our trials in Figure 5.23 in order to demonstrate this relationship.

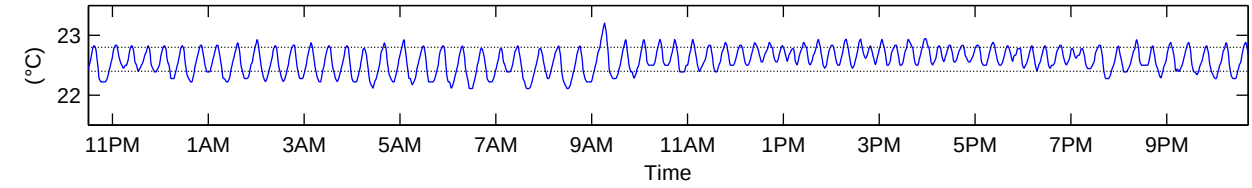
To produce the results above, we executed a total of 330 simulation runs of the classic DR model in three separate trials. We adjusted the length of the DR event from 1 to 6 hours in half-hour increments and did this 10 times in each trial for a total of 110 runs per trial. In each trial we picked large values for N_0 ($N_0 = 30$, 40, and 50 for trials 1, 2, and 3, respectively) since DR events typically occur in times of high demand. And so each curve in Figure 5.23 is the mean curtailment ratio over 110 runs in each of these trials. We see that the depth of curtailment over varying DR event durations provides a proxy for quantifying

the deferrability of the loads we are modeling without directly knowing the energy slack of these loads. Therefore, the deferrability of the system in this scenario varies indirectly with the length of a DR event.

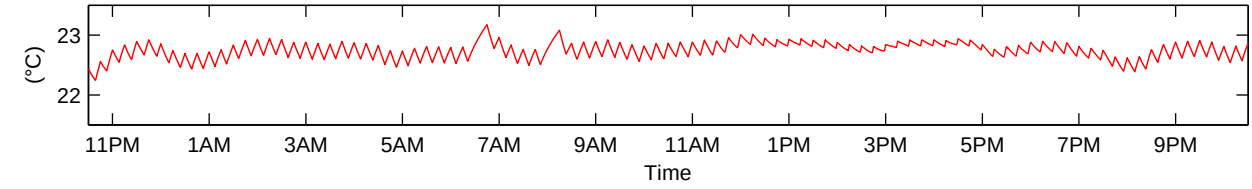
We then assess the performance of our algorithm tailored to the continuous DR scenario using public solar PV data from rooftops across 12 locations around the United States for 5 days of summer (7/8/2010-7/12/2010) and 5 days of winter (12/19/2010-12/23/2010) for a total of 120 day-location pairs of solar power traces [110]. Each day-location pair corresponds to a unique simulation run, all of which are then used to build performance trials based on different supply prediction techniques.

In these trials we wish to determine how much reliance on external power supply (*e.g.*, in the form of utility-provided electricity, on-site energy storage, or backup generation) is needed to buffer demand as the renewable supply fluctuates. To quantify this fluctuation, we invoke the notion of *scaled incremental mean volatility* (scaled IMV) [107] which is simply the absolute difference between the moving average of the supply signal and the actual observed supply, averaged over the time domain of interest and scaled according the ratio of total energy demanded over total energy supplied in that domain. The scaled IMV gives us insights into the high-frequency fluctuations of our supply signal which could be the result of sporadic cloud cover (in the case of solar) or microbursts (in the case of wind) without bias to energy imbalances in supply and demand.

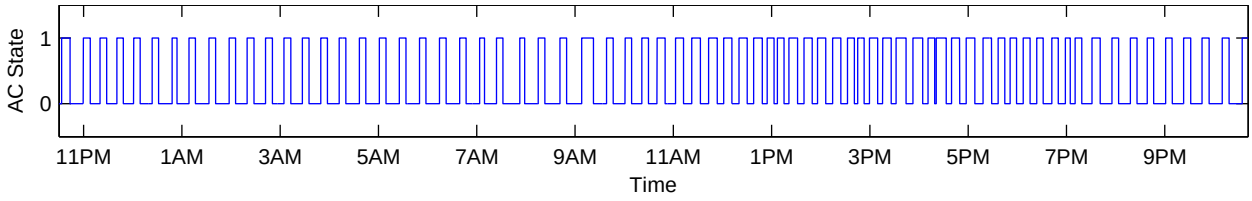
In Figure 5.24 we present the results of four distinct trials corresponding to one unmanaged base case scenario in which our algorithm is not applied, and three applied scenarios employing three distinct supply prediction models for performance comparison. The base case (solid black) assumes loads obliviously consume power regardless of the renewable supply, in which case we do not apply any load control. In one algorithm-applied trial we use a persistence model (dotted red) in which we assume that the power supply in the current time step will be the same in the next time step. In the next trial we use a 1-hour moving window average (dashed green) of historical data to predict supply in the next time step. In the final trial we use an oracle (dash-dot blue) in which we know exactly how much solar will be available in every time step. The average base case grid dependence of oblivious charging is 0.561. The average grid dependencies over each algorithm trial of 0.411, 0.380, and 0.371, respectively, reflect a 26.8-33.8% improvement to the base case when our sort-and-assign knapsack algorithms are applied to laptop charging schedules.



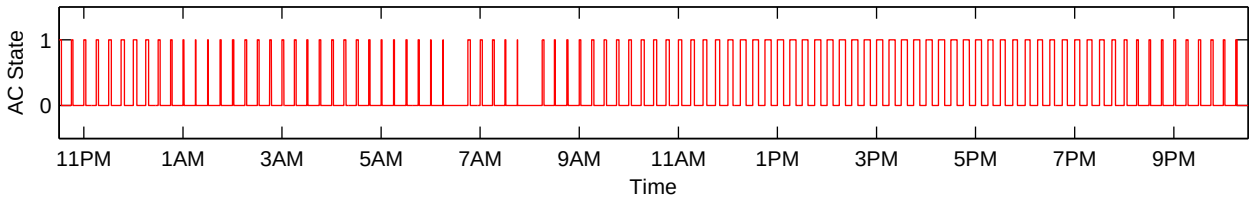
(a) Experimental Two-Position Temperature



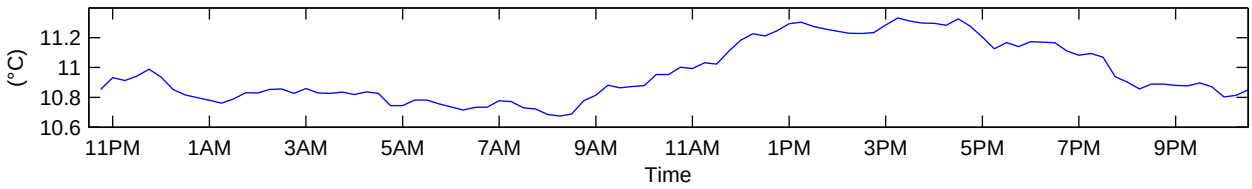
(b) Simulated Learning-Based MPC Temperature



(c) Experimental Two-Position PWM

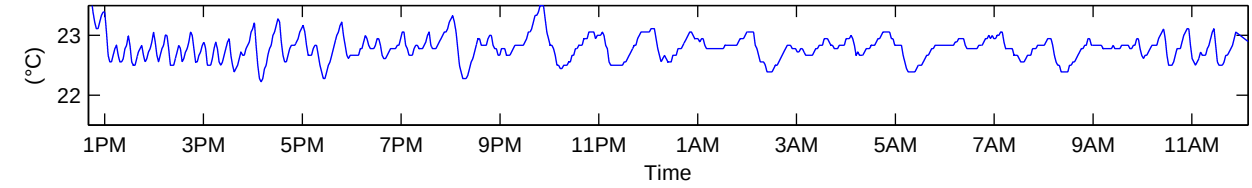


(d) Simulated Learning-Based MPC PWM

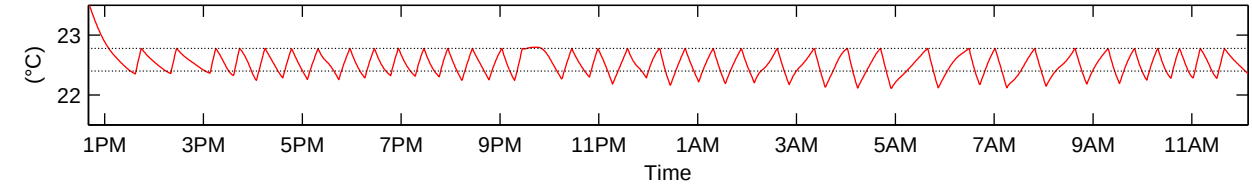


(e) External Heating Load

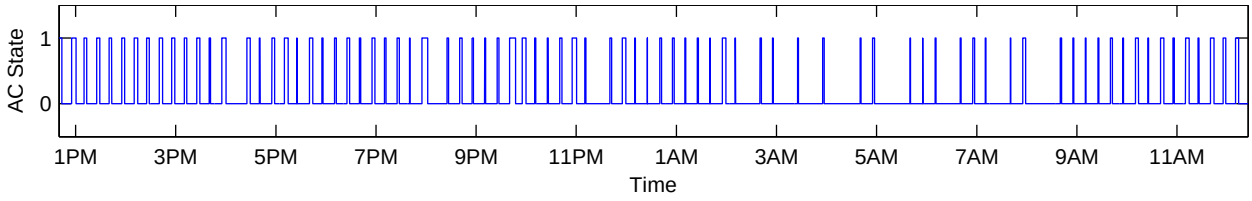
Figure 5.13: The AC was controlled by the two-position control of the thermostat, and the corresponding measured room temperature is shown in units of ($^{\circ}C$). A simulation of the learning-based MPC is given in ($^{\circ}C$). The two-position control uses $32.6kWh$ (estimated $35.1kWh$) of electrical energy, and the learning-based MPC is estimated to use $23.6kWh$. The PWM control generated by the two-position and learning-based MPC control are also shown. An AC state of 0 corresponds to the AC off, and AC state of 1 corresponds to the AC on. The external heating load over 15 minutes due to weather and occupancy $k_w w[n] + q[n]$ is given in ($^{\circ}C$).



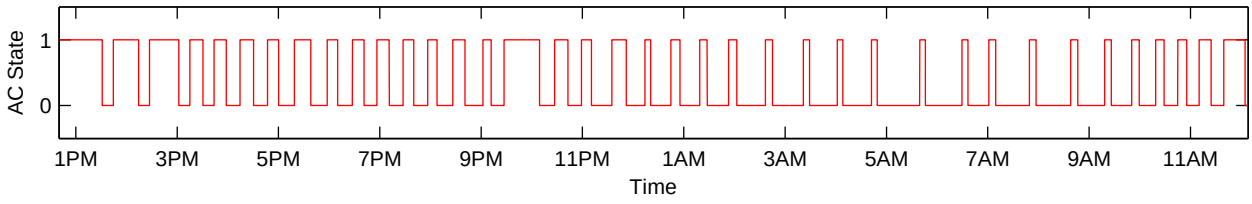
(a) Experimental Learning-Based MPC Temperature



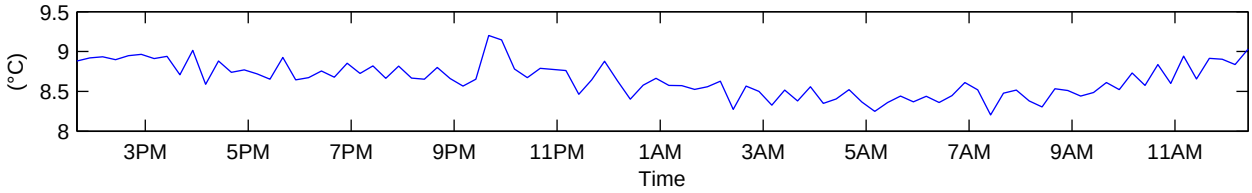
(b) Simulated Two-Position Temperature



(c) Experimental Learning-Based MPC PWM



(d) Simulated Two-Position PWM



(e) External Heating Load

Figure 5.14: The AC was controlled by the learning-based MPC, and the corresponding measured room temperature is shown in units of ($^{\circ}C$). A simulation of the two-position control is given in ($^{\circ}C$). The learning-based MPC uses $11.8kWh$ (estimated $13.3kWh$) of electrical energy, and the two-position control is estimated to use $34.5kWh$. The PWM control generated by the learning-based MPC and the two-position control are also shown. An AC state of 0 corresponds to the AC off, and AC state of 1 corresponds to the AC on. The change in temperature over 15 minutes corresponding to experimentally measured weather and occupancy $k_w w[n] + q[n]$ is provided in ($^{\circ}C$).

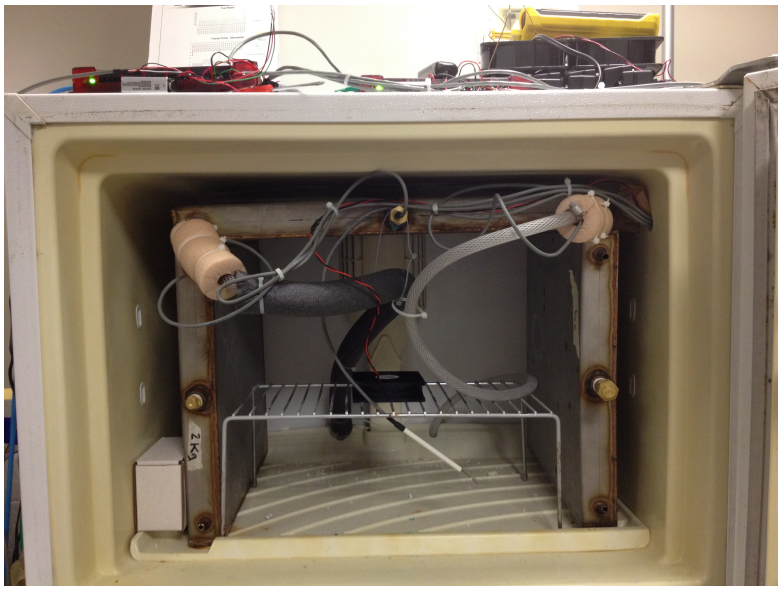


Figure 5.15: Picture of the experimental setup, showing freezer contents. External pump and heat exchanger not shown.

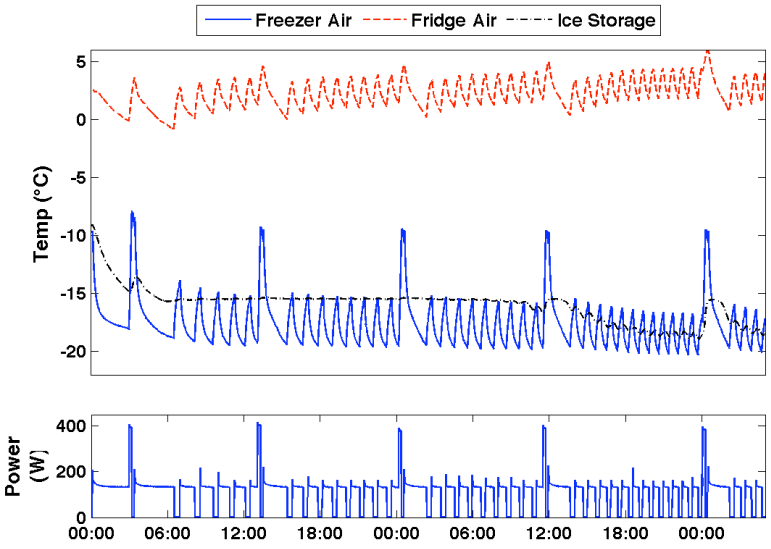


Figure 5.16: Operation of energy storage fridge. Note the thermal storage element charging as the material becomes fully frozen and enters steady state.

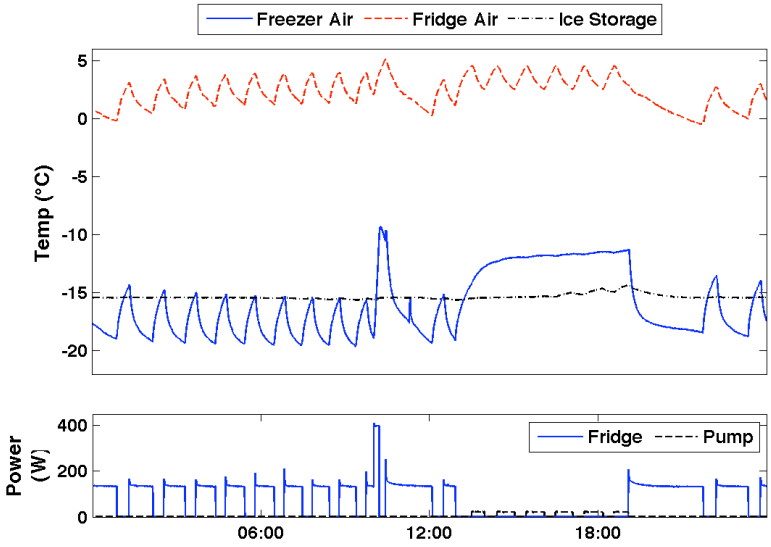


Figure 5.17: Operation of energy storage fridge using a time-of-use price schedule.

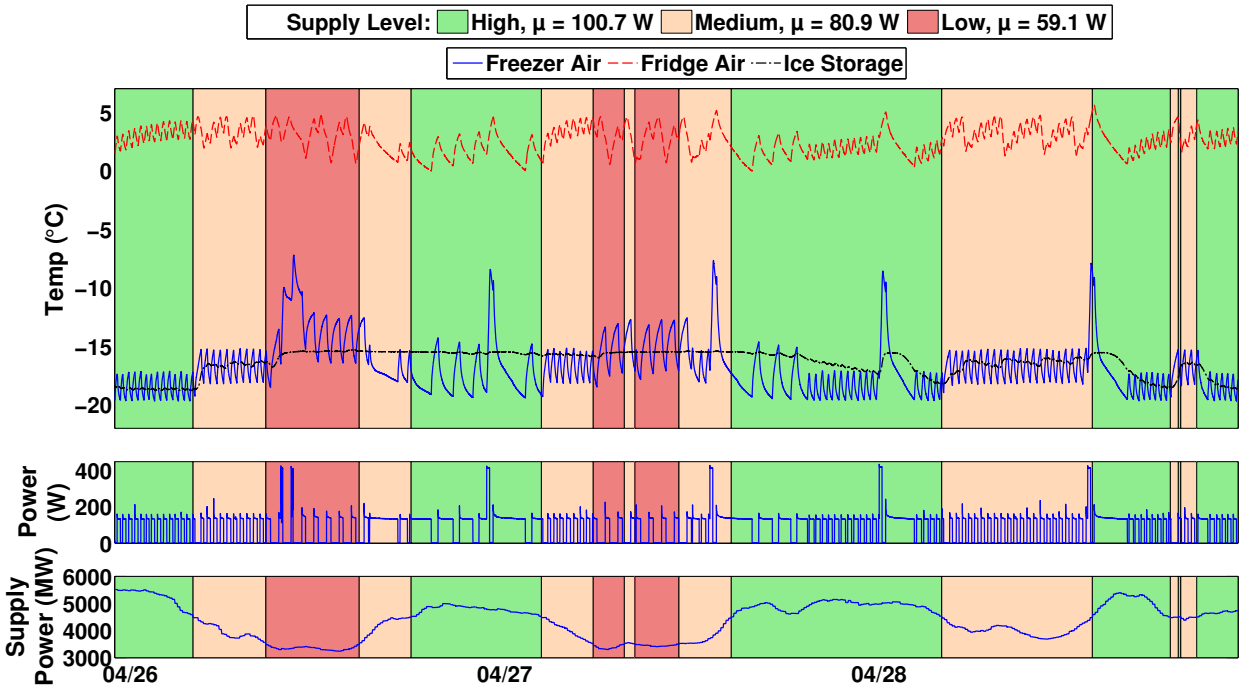


Figure 5.18: Operation of the thermal storage refrigerator as a supply-following load. The refrigerator is configured to work in three modes – High, Medium, and Low – based on the availability of renewable power using a real-time signal from CAISO.

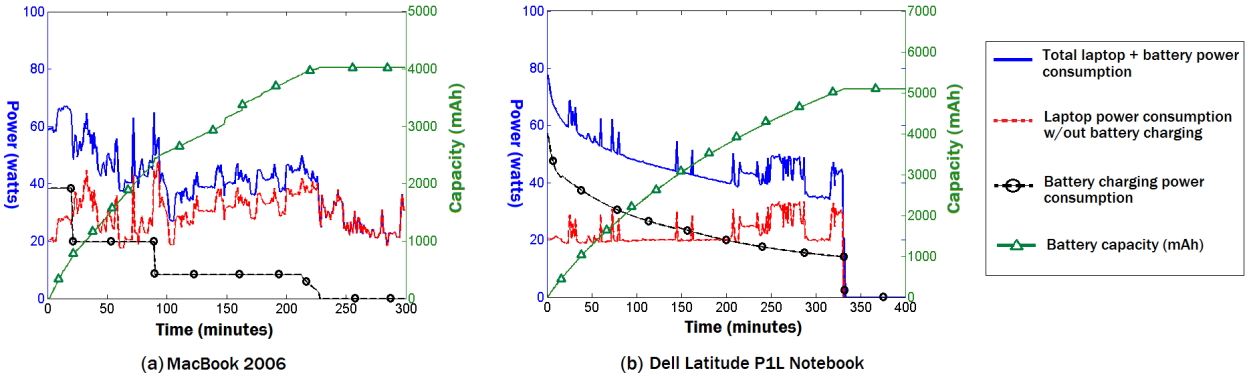


Figure 5.19: On the left we observe one full charging cycle of a 2006 Apple MacBook, and on the right we observe a Dell Latitude P1L. The solid blue curve in each graph represents the total power going into each laptop and their batteries. The dashed red curve represents the power delivered to just the laptops, and the circle-dashed black curve is the power delivered to their batteries alone; and thus, the sum of the circle-dashed black and dashed red curves equal the solid blue curve in each figure. The triangle green curves depict the battery capacity of each laptop varying over time as they charge from 0% to 100%.

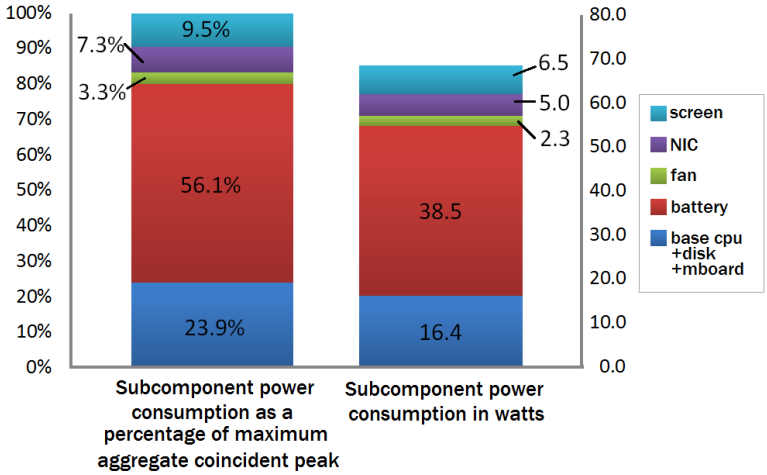


Figure 5.20: Coincident subcomponent power disaggregation of an Apple MacBook 2006.

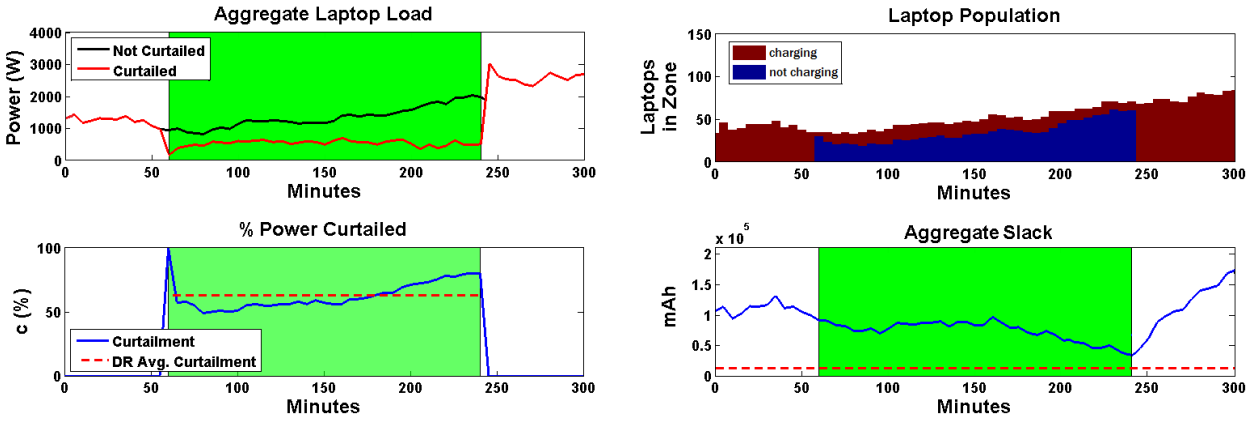


Figure 5.21: Simulation run of initially 50 laptops in 5-minute time steps over a total of 5 hours interrupted by a 3-hour DR event (green shade) that begins at $t = 60$min and ends at $t = 240$min. The dashed red line in the slack figure represents the sum of acceptance thresholds of the initial 50 laptops. As the total slack of the load approaches this aggregate acceptance threshold, the deferrability of the entire system approaches zero.

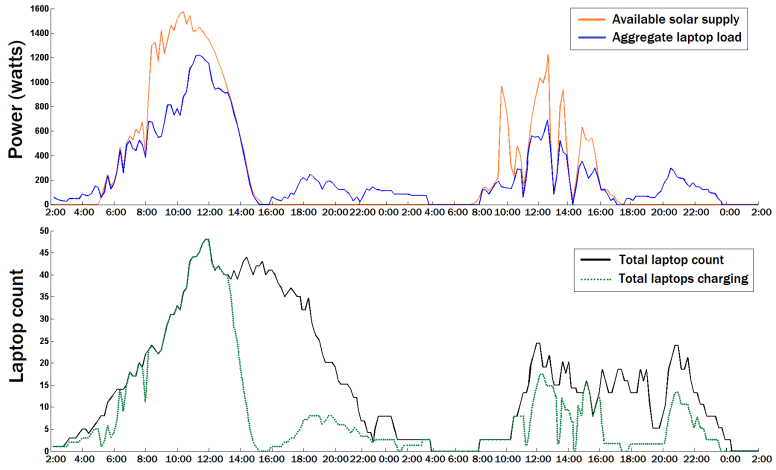


Figure 5.22: Simulation run of continuous DR over two days in 15 minute time steps.

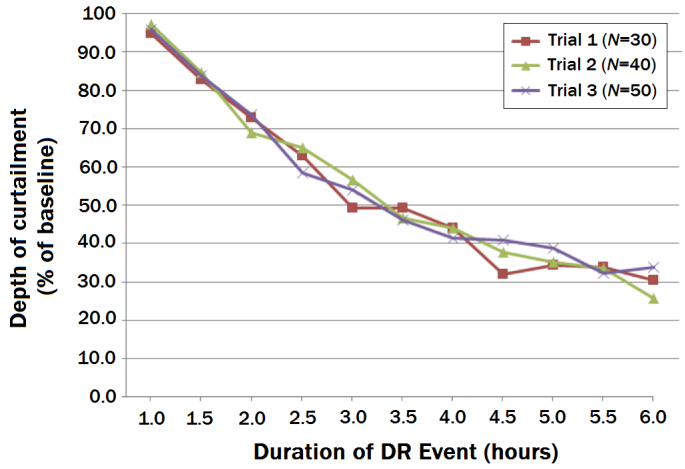


Figure 5.23: Depth of curtailment as percent of baseline load with respect to DR event duration.

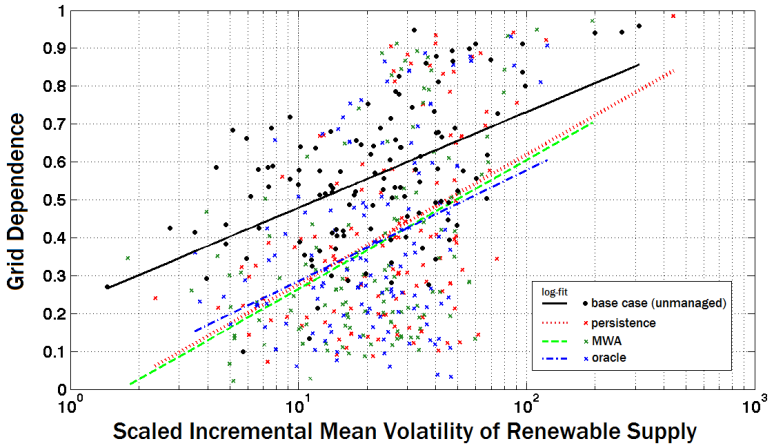


Figure 5.24: Semilogarithmic scatter plot of grid dependence as a function of the scaled IMV of the supply signal for a specific day (*i.e.* simulation run) of all three supply prediction trials compared to the unmanaged base case trial without any charge scheduling algorithms applied.

Chapter 6

Supply-Following Loads at Scale

In Chapter 5, we presented and evaluated a set of implementations of supply-following electricity loads. From these concrete examples, we can analyze the capacity for individual loads to provide energy flexibility in isolation. In this chapter, we scale this capacity by considering collections of supply-following loads, apply the techniques presented in this work cooperatively, and assess the remaining challenges of a grid with deep renewables penetration and a significant population of supply-following electricity loads.

First, we examine the behavior of a collection of the supply-following refrigerators introduced in Section 5.2.2 to understand an important result of aggregation of thermostatically-controlled supply-following loads. Next, we incorporate the measured behavior of the refrigerator enhanced with thermal storage from Section 5.4.1 into a model of a supply-following load. We couple collections of this model with models of the California electricity grid at various levels of renewables penetration created using the methodology in Chapter 3; this allows evaluation of the effects of different levels of penetration of supply-following technology into the refrigerator fleet in California at different levels of renewables penetration. This analysis exhibits the potential benefits of incorporating supply-following load technology into the grid as renewables penetration increases. We then adapt observations from the implementation of an efficient HVAC controller in Section 5.3 and from other work [129] to quantify the flexibility potential of building HVAC systems; we then use this model to evaluate the supply-following potential of this type of load on the California grid. Finally, we bring together these grid interventions with the hydroelectric management strategy presented in Section 4.1 to evaluate the overall effect of incorporating large populations of supply-following loads, assessing the improvement in the match between supply and demand in a grid with deep renewables penetration. From this, we can assess the remaining variability, laying out an agenda for further research.

6.1 Collections of Supply-Following Loads

In Chapter 5, we implemented and characterized the behavior of individual supply-following loads. Building on that, in this section we examine properties of collections of supply-following loads, first as an aggregate of many individual loads and then in the context of an electricity grid. Our experiments in this section employ the two types of refrigerators profiled in this work: supply-following refrigerators and refrigerators with enhanced thermal storage.

6.1.1 Supply-Following Refrigerators

We first consider the effects of aggregating the responses of a population of supply-following loads. For this experiment, we use the model of a supply-following refrigerator without enhanced thermal storage described in Section 5.2.2. In Figure 6.1, we show the proportional change between renewable and portfolio energy as the aggregated population of loads increases. In this scenario, each fridge is initialized at a random phase with warming, cooling, and electrical behavior randomly selected from the normal distribution described in Section 5.2.1.1. Additionally, fridge responses are scaled to the magnitude of the energy deficits and surpluses, just as in the "wide guardband" heater described in Section 5.2.1.1. Further – to more accurately represent actual behavior due to measurement error and local microclimatic effects – a small amount of randomness has been applied to the guardband boundaries, such that each time a boundary is approached, its value is increased or decreased by a random value within +/- 0.2°C. This small modification, while not capturing the entire variation among a large population of appliances, is beneficial in reducing the "herd effects" when a number of identical agents with identical control logic respond to the same stimulus. We believe that in practice, a population of fridges will in fact exhibit more variation than represented in this experiment due to variations in factors such as indoor and outdoor weather conditions, usage patterns, and age of equipment. Also in this experiment, the wind energy supplied is scaled to satisfy 25% of the energy needed by the "oblivious" fridge. This restriction highlights the benefits of statistical multiplexing in a population.

Looking at the results of the population study, the proportion of total energy supplied by renewables used by the fridges improves rapidly as the number of fridges increases, leveling off around 85% at a modest level of 25 fridges. With such limited wind energy supplied per fridge – with only one fridge, the wind supplies barely over 25% of the energy required for operation – there is significant improvement as each fridge is added. This is because fridges with slightly different guardbands and varied warming and cooling behavior operate at slightly different periods and phases, allowing individual fridges to use excess renewable energy not used by others in the population. As the population grows, the collection of power spikes resulting from the compression phases of each fridge blend into a flatter curve that begins to resemble the wind supply curve, reducing the overall difference between the consumption and the supply. This statistical multiplexing of a collection of intermittently operating devices, as many thermostatically-controlled are, shows the enormous potential of

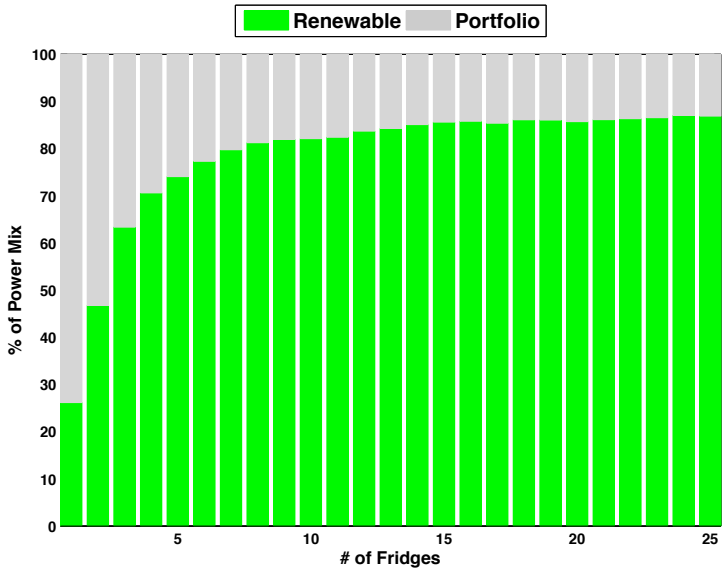


Figure 6.1: The effects of population size on the proportion of total energy that is renewable. Variations of fridge periods and phases allows for cooperation without direct communication, making better use of renewable resources.

supply-following loads to consume fluctuating supplies of power. This result is only enhanced by the likely amplified variations among an actual population of fridges.

6.1.2 Refrigerators with Enhanced Thermal Storage

As shown, a collection of non-thermally enhanced refrigerators has substantial agility and is able to nearly match a fluctuating supply of wind energy. Nonetheless, in Section 5.4.1, we make the observation that the flexibility of this load would be further improved by increasing the available energy storage. In this section, we seek to quantify the grid-scale effects of collections of supply-following loads. To do this, we employ the two applications for the refrigerator with enhanced thermal storage that were introduced in Section 5.4.1, price-responsive demand and supply-following. We look at these applications at grid-scale, using the model developed in Chapter 3 for for the California electricity grid at deep renewables penetration.

6.1.2.1 Price-Responsive Demand

To quantify the potential of flexible capacity, we provide data about the population of refrigerators currently owned in California in Table 6.1. We use these numbers as a guide to produce a bottom-up estimate of the total refrigeration and freezer load in the state, nearly 2.6 GW. Additionally, the California and federal governments conducted periodic large-scale

surveys to estimate end-use consumptions; these numbers can be used to make a top-down estimate of power consumption of this end use. Refrigeration and freezing represent 14.8% of total commercial and 20% of total residential electricity consumption in California, sectors which represent 47% and 35% of electricity in the state, respectively [23, 24]. Combining these numbers indicates that 14% of total electricity consumption in the state of California, or roughly 3.6 GW on average, is for residential and commercial refrigeration and freezing, composed of units ranging from small to warehouse-sized. For our model, we assume a median point of 3.1 GW total consumption. Using this estimate, we create a model of flexible refrigerator load for the state. Informed by our prototype, the flexible refrigerator load can turn off for up to six hours a day, five days a week, provided it can charge the rest of the week by drawing additional power (13% extra power due to the storage element).

To study the applications of this flexible load at grid-scale, we employ our methodology for scaling renewables penetration in electricity grids described in Chapter 3, examining the effects of a wide deployment of flexible refrigerators on the California electricity grid at a range of levels of renewables penetration. We also vary the depth of adoption of this technology. We first evaluate the grid-scale effects of the price-responsive demand application. Since many utilities are regulated, their ability to price electricity to reflect market demand is limited; often, during periods of high demand and high supply costs, they cannot recoup their expenses with higher rates. Thus, utilities are often incentivized to get their customers to avoid power consumption at peak hours, in the short term to avoid operating expensive, environmentally-unfriendly peaker electricity plants, and in the long term to avoid having to provision and maintain such facilities. Thus, TOU pricing tariffs are a lever for utilities to reduce the high costs of providing peak power by flattening the peak. Table 6.2 shows the effect on the peak total demand of the California grid over a year as a series of load duration curves, providing a statistical assessment of the distribution. Each curve represents a different rate of adoption of price-responsive flexible refrigerator technology. As we are observing the level of peak demand, the generation mix of the grid is not considered. The fridges operate on a time-of-use schedule, avoiding consumption from 1-7 PM every week day and charging at all other times. The data show the massive benefit of a flexible load to grid operators – since high demand is often correlated with periods of peak pricing, the benefits of flexible capacity can directly reduce peak generation capacity needed. Though the mean of demand is also slightly reduced, the effect on the more expensive peak is far more dramatic.

6.1.2.2 Supply-Following

Beyond price-responsive demand, grid operators can employ this technology in other ways. For example, a fleet of refrigerators could provide balancing to operate a renewable generation source as if it were a baseload power plant, or flexible refrigerators could flatten duration curves of fossil fuel resources rather than the total demand on the grid, perhaps obviating the need to maintain underutilized power plants. Using the same model of flexible refrigerator load, we allow the supply-following refrigerator to avoid consumption any six hours a day,

	Units [21, 123]	Avg. Power [21, 24]	Total Avg. Load
Residential			
First Unit	9.4 M	88.1 W	0.83 GW
Add'l Units	3.7 M	138.4 W	0.51 GW
Freezers	1.8 M	106.6 W	0.19 GW
Total	14.9 M	102.8 W	1.53 GW
Commercial			
Total	425 k (est.)	2.4 kW	1.03 GW
Overall	15.3 M	166.5 W	2.56 GW

Table 6.1: Statistics about residential and commercial refrigerators and freezers in California.

	Fridge Adoption Rate					
% of Hours	**0%** (GW)	**10%** (GW)	**20%** (GW)	**33%** (GW)	**50%** (GW)	**100%** (GW)
0.0 (Max)	47.128	46.818	46.508	46.095	45.578	44.088
0.1	45.054	44.744	44.434	44.020	43.846	43.235
0.5	42.325	42.014	41.754	41.420	40.999	39.883
1.0	39.755	39.610	39.548	39.382	39.252	38.487
5.0	34.828	34.701	34.621	34.504	34.309	33.988
10.0	32.069	31.972	31.944	31.843	31.751	31.573
50.0 (Median)	26.173	26.108	26.015	25.898	25.722	25.373
100.0 (Min)	18.792	18.832	18.873	18.926	18.994	19.195
Mean	26.314	26.291	26.269	26.239	26.202	26.091

Table 6.2: Summaries of duration curves for total supply on the California grid. Columns represent adoption rates of thermal storage refrigerators in California operating on time-of-use tariffs, avoiding consumption 1-7 PM on weekdays.

and configure the controller to select the hours with the most natural gas generation.

Table 6.3 provides the peak natural gas generation required for the year at different levels of flexible refrigerator adoption and renewables penetration. Again, the fraction of flexible fridge capacity is able to effectively target the worst hours, reducing peak use of natural gas generation by the same capacity. Further, as the penetration of renewables increases, the percent reduction in peak grows, as the peak need for natural gas has also been reduced by the larger deployment of renewables. Though not shown, these benefits are concentrated at the top of the duration curve, as the mean of the natural gas resource is not substantially altered. Working in concert, flexible loads become more valuable as penetration of renewables increases. Though preliminary, we believe these results are promising for the potential of this important, emerging class of flexible electric loads.

6.1.3 Flexibility in Commercial Building HVAC Schedules

Another potential target for becoming supply-following is heating, ventilation, and air conditioning (HVAC) systems. As stated in Section 5.3.1, HVAC accounts for roughly a third of building energy usage. Our implementation of an energy-efficient HVAC controller shows that there is enormous potential to improve operation of existing HVAC equipment through

Peak Natural Gas Supply (GW)						
	Fridge Adoption Rate					
Renewables Penetration (%)	**0%** (GW)	**10%** (GW)	**20%** (GW)	**33%** (GW)	**50%** (GW)	**100%** (GW)
Current (11%)	27.014	26.704	26.394	25.981	25.464	23.914
20%	27.014	26.704	26.394	25.981	25.464	23.914
30%	27.014	26.704	26.394	25.981	25.464	23.914
40%	26.496	26.186	25.876	25.463	24.946	23.396
50%	23.634	23.325	23.015	22.601	22.084	20.535
60%	19.528	19.218	18.908	18.495	17.978	16.428
70%	17.057	16.747	16.437	16.024	15.507	13.957

Table 6.3: Peak natural gas generation on the California grid. Columns represent adoption rates of supply-following thermal storage refrigerators in California. Rows represent levels of renewables penetration on the grid.

improved control techniques; those experiments showed systems that were anywhere from 30% to 70% more efficient.

To scale those results to our grid model, we employ a study that describes a whole building demand response system for commercial buildings [129]. In this study, the authors created a series of models for buildings, varying the building size (small, medium, and large), material used (concrete and steel), climate zone (warm, hot, and extremely hot), thermal mass level (light, medium, and heavy), and demand response (DR) strategy used (linear, step, and exponential), for a total of 162 different building scenarios. DR strategies were implemented with a precooling stage, where the building was cooled beyond normal levels, consuming additional power in advance of a known DR event, followed by a reduction stage, where the building was allowed to become warmer than normal levels, greatly reducing power. Using EnergyPlus [121], a building energy simulator, the authors derived results of potential for demand response for each of these buildings, providing a percentage of whole building power that could be shed in the reduction stage for a period of time. The results of this work showed that building types generally increased power consumption by 8-10% for the precooling stage before reducing whole building power by 12-24% for the reduction stage, depending on building material type, size, thermal mass, DR strategy, and climate zone. Using these results, we defined a simplified model of demand response using thermal mass that aimed to be indicative of the population of commercial buildings in California. Using representative values from the study, our model uses a four-hour increase in whole building power by 8% for the precooling stage followed by a six-hour decrease in whole building power by 18% for the reduction stage.

To calculate the size of the potential for this technique on the California grid, we employed statewide data [22] that indicate 38% of electricity consumption is in the commercial sector. Using the California End-Use Survey [21], we calculate that 24% of this portion goes to office buildings. With these two proportions in tandem with our California grid model from Chapter 3, we can now estimate the grid-scale effects of employing thermal mass-based demand response strategies on the office building stock in California.

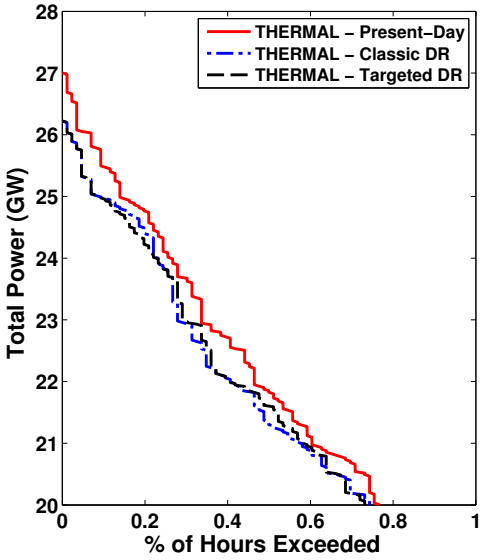
We employed two different demand response strategies: classic DR, as introduced in Section 5.4.2, and targeted DR. Recall that in classic DR, the goal is to reduce peak demand by some percentage of baseline power consumption during a scheduled DR event that is continuous for some known duration. In this case, the duration is always from 12-6 PM, the peak hours of the current California electricity grid. In targeted DR, we employ supply-following, whereby the four precooling and six reduction hours can be opportunistically placed throughout the day to best match supply availability.

We seek to highlight the differences in these strategies as the grid changes from the current state, with 11% renewables, to a future grid with deep renewables penetration (in this case, 60% renewables). Specifically, we examine the leftmost extreme of the duration curve of the thermal (natural gas) resource, the hours where the largest capacity of fossil fuel generation is required. Figures 6.2 and 6.3 show this portion, the top 1% of the duration curve of the thermal resource, for the current-day and scaled to 60% renewables grids. Using the tables included with the figures, we can see that in the current-day grid the classic DR approach performs similarly to the targeted DR approach, reducing the peak thermal capacity requirement by 0.780 GW. Differences between the two approaches are minimal through the rest of the duration curve. This is because the targeted DR approach tends to choose the same hours as the classic DR approach, since peak natural gas periods tend to correlate with the peak demand hours from 12-6 PM. In the case of the grid scaled to 60%, we can see that the targeted DR approach now clearly outperforms the classic DR approach throughout the duration curve. This is because peak natural gas hours no longer occur during the 12-6 PM period. This finding highlights the necessity of information flow to loads when trying to reduce emissions in future electricity grids; without knowing when peaks in the natural gas generation occur, naïve approaches that work only using the hour of the day are no longer effective at reducing emissions.

6.2 Assessing Remaining Variability

A natural question that arises from this study is to what extent supply-following loads can address the mismatch between supply and demand in grids with deep penetration of renewables generation. In this section, we address that question by combining the grid-scale supply-following techniques introduced in this work, using our model of the California electricity grid scaled to 60% renewables penetration. Specifically, we employ the hydroelectric management strategy from Section 4.1, the building thermal mass demand response from Section 6.1.3, and the refrigerators enhanced with thermal storage from Section 5.4.1. Using these three strategies, we assess the overall effect of a large population of supply-following loads on a deeply renewable electricity grid, with particular interest in quantifying the remaining discrepancies between supply and demand (*i.e.*, the remaining need for thermal generation) after applying supply-following loads.

Figure 6.4 shows two timeseries of the 60% renewable electricity grid: first with no additional modification to the scaled grid, and second by applying the hydroelectric management,

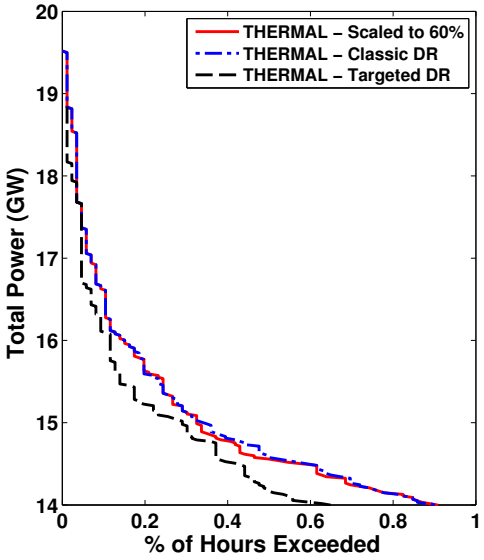


%Hrs	GW	GW	GW
Max	27.014	26.234	26.234
0.1	25.779	25.016	25.016
0.5	21.887	21.364	21.622
1.0	19.144	18.562	18.540
5.0	14.812	14.613	14.559
10.0	13.414	13.284	13.236
50.0	8.368	8.328	8.319
100.0	2.720	2.576	2.720

Figure 6.2: Thermal duration curves for the top 1% of hours in the present-day CAISO grid. Three scenarios are shown: an unmodified grid, a grid employing classic demand response for commercial buildings (reduction from 12-6 PM), and a grid employing targeted demand response for commercial buildings (reductions on the six hours with the highest thermal generation).

building thermal mass DR, and thermal storage refrigerators, in that order. This order was selected to go from the most centralized approach, hydroelectric management, to the most distributed approach, thermal storage refrigerators. In reality, these systems would all be reacting to changes from each other, but for this implementation, it makes the most sense to have the highly distributed refrigerators that have the highest resolution respond to the more centralized responses with the least resolution. From the time series, we can see that the gains are quite modest, as less than one percent of overall energy consumption shifts from non-renewable to renewable generation. Figures 6.5 and 6.6 look more closely at five days in the summer and winter in this grid. In the summer, we can see that excess dominates, so there is little benefit from gains via any of the shifting strategies. In the winter, we see brief periods of excess, so shifting strategies like the thermal mass building demand response and the supply-following refrigerators do little to change the overall balance of resources consumed. In fact, the percentage of electricity delivered from fossil fuels marginally increases during these days because of the increased consumption of the supply-following refrigerators. The hydroelectric management strategy has some positive effect, but it is minimal because of the limited availability of this resource during the winter season due to the lack of excess generation.

The real benefit of these supply-following loads is from the improvement at the leftmost edge of the duration curve of the thermal resource. Figure 6.7 exhibits this duration curve in

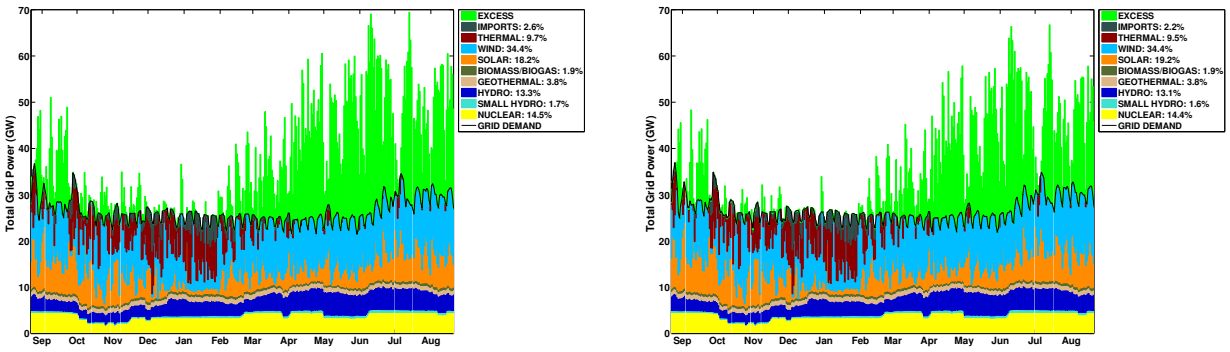


%Hrs	GW	GW	GW
Max	19.528	19.528	18.878
0.1	16.693	16.693	16.313
0.5	14.569	14.618	14.212
1.0	13.905	13.885	13.562
5.0	11.314	11.279	11.128
10.0	9.588	9.520	9.465
50.0	0.000	0.000	0.000
100.0	0.000	0.000	0.000

Figure 6.3: Thermal duration curves for the top 1% of hours in the CAISO grid scaled to 60% renewables. Three scenarios are shown: a scaled but otherwise unmodified grid, a scaled grid employing classic demand response for commercial buildings (reduction from 12-6 PM), and a scaled grid employing targeted demand response for commercial buildings (reductions on the six hours with the highest thermal generation).

two ways. First, in Figure 6.7(a), the entire duration curve is shown, and in Figure 6.7(b), the view is zoomed to the top 1% of hours of natural gas generation. From this figure, we can see that during the top 1% of hours, the hydroelectric management strategy has essentially zero effect at the most critical hours of the year; this is because the hydroelectric resource during this portion of the year tends to be near its minimum. As other strategies are applied, the effect at the peak is more visible – the peak hour is reduced from 19.5 GW to 15.8 GW, a reduction of 19%. We also see the strong flattening effect the thermal storage refrigerators have on the duration curve. The peak hours are reduced so that the highest natural gas generation required occurs in a number of different hours, rather than in a single hour; this gives the duration curve a stairstep pattern. Further, we can see that generation has been relocated to other parts of the duration curve – the generation required for more than 95% of hours either stays the same or increases after supply-following refrigerators are included. As designed, these are not the critical hours, so this shows how supply-following strategies can enormously benefit the grid.

After applying supply-following techniques to reduce peak natural gas capacity by 19%, we end up with a massively different grid. Though there are still seasonal imbalances in the availability of renewables, the fleet of generation resources needed to provide electricity in such a grid is substantially different. By reducing the sharp peak of the natural gas generation duration curve, the scope of potential improvements has broadened, allowing a



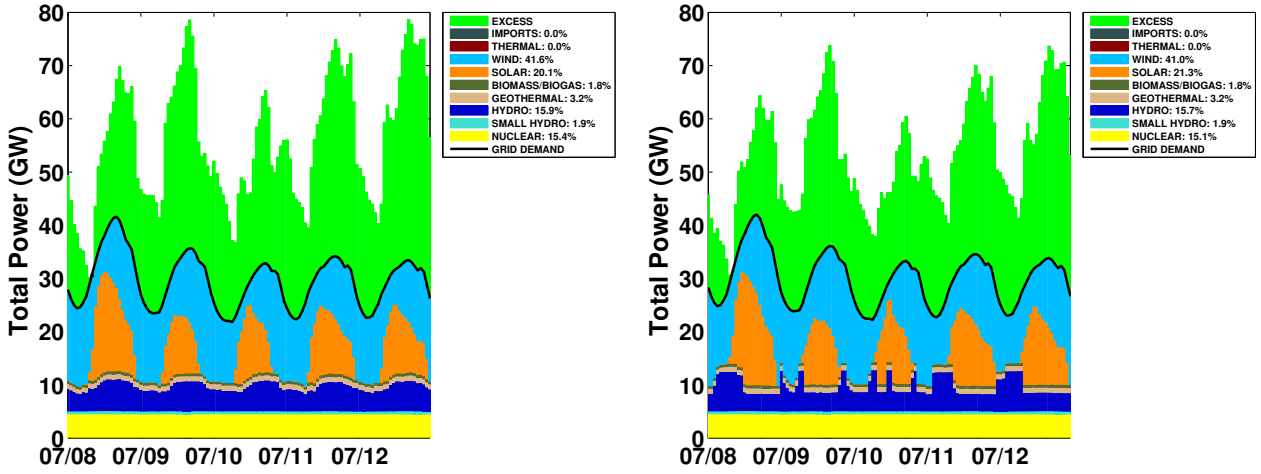
(a) Scaled grid with no modifications

(b) Scaled grid with Hydro Management, Building Thermal Mass DR, and Thermal Storage Fridges

Figure 6.4: California electricity blend for a year, scaled to 60% renewables in scenarios before and after applying supply-following strategies.

far greater range of strategies to have an impactful effect on supply and demand matching in the grid.

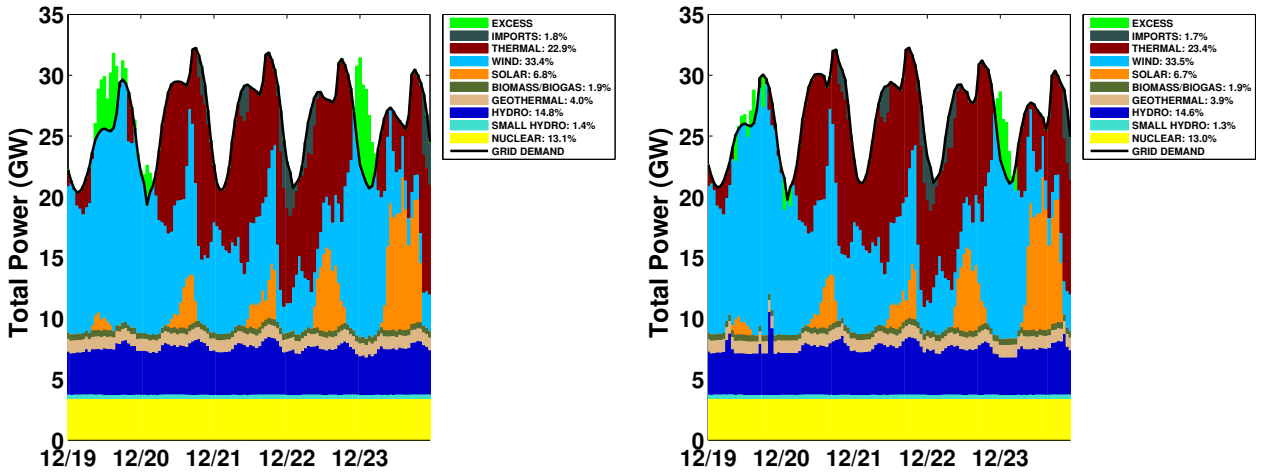
With a better understanding of what the challenges of a future grid might look like, we can begin to outline a research agenda for preemptively solving the critical problems that may occur. Specific improvements include but are not limited to improved efficiency of winter and nighttime loads (*e.g.*, heating and exterior lighting), further deployment of solar-thermal generation that can delay generation later into the evening, increasingly flexible fuel-based generation sources that are able to ramp up and down more quickly, and seasonal energy storage that can shift multiple gigawatts of generation from summer to winter. These represent advancements that become much more valuable in grids with deep renewables penetration, as they either relieve the newly uncovered peaks in fossil fuel-based generation or they allow solutions to the broader challenges that are now accessible since the need to address the peak is less severe.



(a) Summer - Scaled grid with no modifications

(b) Summer - Scaled grid with Hydro Management, Building Thermal Mass DR, and Thermal Storage Fridges

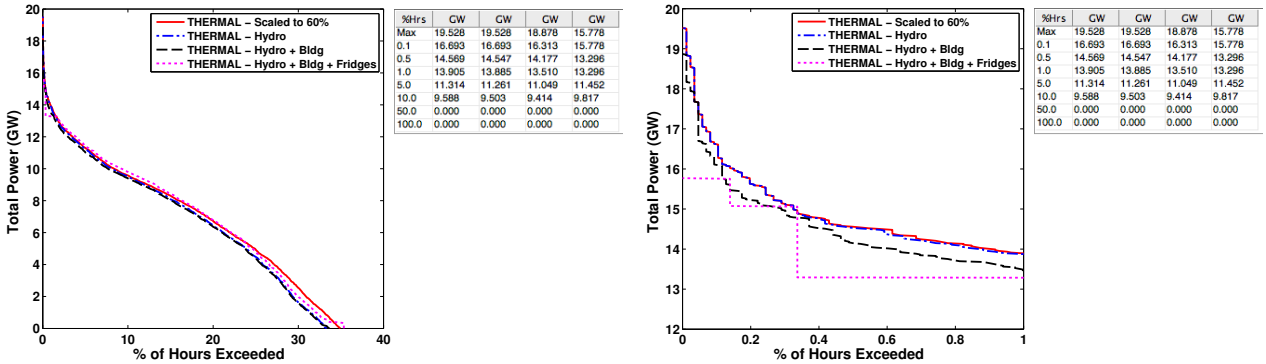
Figure 6.5: Five summer days of California electricity blend scaled to 60% renewables (a) with no further modification and (b) with hydroelectric management, building thermal mass demand response, and thermal storage fridges.



(a) Winter - Scaled grid with no modifications

(b) Winter - Scaled grid with Hydro Management, Building Thermal Mass DR, and Thermal Storage Fridges

Figure 6.6: Five winter days of California electricity blend scaled to 60% renewables (a) with no further modification and (b) with hydroelectric management, building thermal mass demand response, and thermal storage fridges.



(a) Duration curve of scaled grid with Hydro Management, Building Thermal Mass DR, and Thermal Storage Fridges

(b) Same duration curve, zoomed in to peak

Figure 6.7: Duration curves for the Thermal resource on the California grid for a year, scaled to 60% renewables in scenarios before and after applying supply-following strategies.

Chapter 7

Conclusion

While electric grids were originally designed in an era when energy was inexpensive and plentiful and communication was almost non-existent, they are being transformed in an age of zero-emissions production and pervasive information. Modern grids present a series of challenges that demand a rich information plane and distributed intelligence: the need to orchestrate a heterogenous collection of distributed resources in the presence of deep uncertainty, where the main elements of the system are actuated by forces outside the domain of control. As we move toward sustainable grids, the uncertainty and underactuation is not just in the energy demand, but also in the generation of renewable energy. This setting defines a completely new set of challenges in order to achieve reliable electricity grid operation.

In this study, we follow a methodology reflecting a mantra of 3Ms: Monitor, Model, and Mitigate. Utilizing large-scale and detailed monitoring of generation resources on three different electricity grids, we are able to represent and characterize the temporal dynamics of grids today at the scale of the state of California, the province of Quebec, and the country of Germany. Based on these and a need to study grids with much deeper renewables penetration, we construct a methodology for scaling the renewables penetration on these grids far beyond their current level while maintaining the temporal and geographical patterns inherent in renewables generation. These models enable us to study the entirely new set of challenges posed by grids with deep renewables penetration, which differ widely from those challenges dominating the discussion today. Rather than critical peak demand driven by cooling loads, the summer has copious energy availability, but the winter has deep lulls. Further, we uncover a surprising observation about the breakdown between solar and wind on these three grids: looking only at energy production, California should have a mix of wind and solar while Ontario and Germany should exclusively employ wind energy. Additionally, we formulate hard and soft limits to renewables penetration, representing thresholds beyond which adding additional renewables capacity to increase renewables penetration is either impossible or so onerous as to be ineffective.

In this new context of a grid with deep renewables penetration, we evaluate the effectiveness of well-studied techniques such as alternative management of existing supply resources, grid-scale energy storage, improved energy efficiency, and demand response. We discover

vastly different roles for these familiar strategies, better serving a grid with deep renewables penetration by targeting the hours of peak natural gas-fueled generation rather than the hours of peak overall demand. With new goals in mind, this provides a fresh agenda for research on these techniques. We also deeply study the role of supply-following electricity loads, which react to supply conditions to shape their electricity consumption patterns. We implement and evaluate a series of these loads, including a typical residential refrigerator and heating unit, an air conditioner used for residential and small commercial buildings, a refrigerator augmented with enhanced thermal storage, and a laptop computer as a proxy for an electric vehicle. These concrete instances provide guidance for models that enable evaluation of the impact of large populations of supply-following loads at the scale of an electricity grid with varying levels of renewables penetration. We show that a combination of supply-following techniques are able to reduce peak natural gas capacity by 19%, an enormous benefit that can substantially alter the economics of operating electricity grids. Beyond this significant improvement, we are able to undertake a clear-eyed assessment of the remaining challenges of supply and demand matching in a grid with large populations of supply-following loads and deep renewables penetration, noting the need for more flexible fuel-based generation and large seasonal stores of energy.

Having explored abstractly several of the critical avenues for mitigating the supply-demand matching problem, we note that each brings benefits yet each has limitations. Some are at the fine scale of rapid ramp rates and coordinated response, while others are at the coarse scale of seasonal variations and total reservoir capacity. To tackle the challenges presented by a sustainable energy grid, all of these techniques and more will need to be employed in a coordinated fashion.

This study is only a very preliminary step, but it makes a significant stride by examining temporal dynamics of a large-scale energy network, rather than only manipulating statistical characterizations of network elements. In this manner, more of the nature of the challenges of operating modern electricity grids is revealed. However, we look only at large-scale interactions and balance at hourly timescales. This analysis needs replication at finer granularity that exposes specific constraints of the transmission network, individual plants, and loads. Furthermore, the algorithms presented are simplistic and merely characterize the potential and limitations of a handful of opportunities. In every respect, these mechanisms and their associated protocols need to be developed far more fully, both to formalize a theory of operation in a sustainable grid and to invent analysis, prediction, and algorithmic techniques that achieve these bounds, as well as technological alternatives that remove such barriers as seasonal limits and ramp rates. Together, these efforts will move us a step closer to the design of a sustainable energy network that eliminates harmful emissions and befits the modern era.

Bibliography

[1] ABB. *GridView*. http://www.abb.com/.

[2] Pieter Abbeel, Adam Coates, and Andrew Ng. "Autonomous Helicopter Aerobatics through Apprenticeship Learning". In: *International Journal of Robotics Research* 29.13 (2010), pp. 1608–1639.

[3] AG Energiebilanzen. *Germany - Electricity Generation by Energy Source from 1990 to 2012*. http://www.ag-energiebilanzen.de/viewpage.php?idpage=65.

[4] Baris Aksanli et al. "Utilizing Green Energy Prediction to Schedule Mixed Batch And Service Jobs in Data Centers". In: *Proceedings of the 4th ACM Workshop on Power-Aware Computing and Systems (HotPower '11)*. 2011.

[5] ANSI/ASHRAE. "Standard 55, Thermal Environmental Conditions for Human Occupancy". In: (2008).

[6] K.J. Åström and B. Wittenmark. *Adaptive control*. Addison-Wesley, 1995.

[7] A. Aswani, P. Bickel, and C. Tomlin. "Regression on Manifolds: Estimation of the Exterior Derivative". In: *Annals of Statistics* 39.1 (2011), pp. 48–81.

[8] A. Aswani et al. "Provably Safe and Robust Learning-Based Model Predictive Control". Submitted.

[9] A. Aswani et al. "Reducing Transient and Steady State Electricity Consumption in HVAC Using Learning-Based Model-Predictive Control". In: *Proceedings of the IEEE* 99 (2011), pp. 1–14. ISSN: 0018-9219. DOI: 10.1109/JPROC.2011.2161242.

[10] Anil Aswani, Peter Bickel, and Claire Tomlin. "Statistics for Sparse, High-Dimensional, and Nonparametric System Identification". In: *International Conference on Robotics and Automation*. 2009.

[11] Anil Aswani et al. "The Berkeley Retrofitting and Inexpensive HVAC Testbed for Energy Efficiency". Submitted.

[12] *Automated Demand Response*. http://www.auto-dr.com/.

[13] J. Barton and D. Infield. "Energy Storage and its Use With Intermittent Renewable Energy". In: *IEEE Transactions on Energy Conversion* 19.2 (2004), pp. 441–448.

[14] P. Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins series in the mathematical sciences. Springer, 1998.

[15] E.Y. Bitar et al. "Bringing Wind Energy to Market". In: *IEEE Transactions on Power Systems* 27.3 (2012), pp. 1225–1235.

[16] F. Borelli, A. Bemporad, and M. Morari. *Constrained Optimal Control and Predictive Control for linear and hybrid systems*. In preparation. 2009.

[17] A. Brooks et al. "Demand Dispatch: Using Real-Time Control of Demand to Help Balance Generation and Load". In: *Power and Energy Magazine, IEEE* 8.3 (2010), pp. 20–29.

[18] *Buildings Energy Data Book*. U.S. DoE, 2009. URL: http://buildingsdatabook.eren.doe.gov/.

[19] W.J. Burke and D.M. Auslander. "Low-frequency pulse width modulation design for HVAC compressors". In: *The ASME 2009 International Design Engineering Technical Conferences*. 2009.

[20] California Council on Science and Technology. *California's Energy Future – The View to 2050*. http://ccst.us/publications/2011/2011energy.php. 2011.

[21] California Energy Commission. *California Commercial End-Use Survey*. http://www.energy.ca.gov/ceus/. 2006.

[22] California Energy Commission. *California Electricity Statistics and Data*. http://energyalmanac.ca.gov/electricity/.

[23] California Energy Commission. *California Energy Consumption by End Use*. http://energyalmanac.ca.gov/electricity/electricity_stats/index.html.

[24] California Energy Commission. *California Residential Appliance Saturation Study*. http://www.energy.ca.gov/appliances/rass/. 2009.

[25] California Energy Commission. *Power Plant Database*. http://energyalmanac.ca.gov/powerplants/Power_Plants.xls.

[26] *California Independent System Operator*. http://www.caiso.com/.

[27] California ISO. *Daily Renewables Watch*. http://www.caiso.com/green/renewableswatch.html.

[28] California Public Utilities Commission. *Demand Response*. http://www.cpuc.ca.gov/PUC/energy/Demand+Response/.

[29] *California Senate Bill 107 - 2006*. http://www.leginfo.ca.gov/pub/05-06/bill/sen/sb_0101-0150/sb_107_bill_20060926_chaptered.pdf.

[30] *California Senate Bill 2 - 2011*. http://www.leginfo.ca.gov/pub/11-12/bill/sen/sb_0001-0050/sbx1_2_bill_20110412_chaptered.pdf.

[31] California Solar Initiative. *California Solar Statistics*. http://www.californiasolarstatistics.ca.gov.

[32] E. Candés, M. Wakin, and S. Boyd. "Enhancing Sparsity by Reweighted L1 Minimization". In: *Journal of Fourier Analysis and Applications* 14 (5 2008), pp. 877–905.

[33] Sheryl Carter, Devra Wang, and Audrey Chang. *The Rosenfeld Effect in California: The Art of Energy Efficiency*. http://www.energy.ca.gov/commissioners/rosenfeld_docs/rosenfeld_effect/presentations/NRDC.pdf.

[34] L. Chisci, J. Rossiter, and G. Zappa. "Systems with presistent disturbances: predictive control with restricted constraints". In: *Automatica* 37 (2001), pp. 1019–1028.

[35] D. Crawley et al. "EnergyPlus: Energy Simulation Program". In: *ASHRAE Journal* 42 (2000), pp. 49–56.

[36] CRC Handbook. *CRC Handbook of Chemistry and Physics, 88th Edition*. 88th ed. CRC Press, 2007.

[37] *Database of State Incentives for Renewables and Efficiency*. http://www.dsireusa.org.

[38] S. Dawson-Haggerty et al. "Enabling Green Building Applications". In: *Sixth Workshop on Hot Topics in Embedded Networked Sensors*. 2010.

[39] S. Dawson-Haggerty et al. "sMAP - a Simple Measurement and Actuation Profile for Physical Information". In: *Proceedings of the Eighth ACM Conference on Embedded Networked Sensor Systems (SenSys '10)*. 2010.

[40] Stephen Dawson-Haggerty et al. "sMAP a Simple Measurement and Actuation Profile for Physical Information". In: *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems (SenSys)*. 2011.

[41] J. DeCesaro and K. Porter. *Wind Energy and Power System Operations: A Review of Wind Integration Studies to Date*. Tech. rep. NREL/SR-550-47256. National Renewable Energy Laboratory, 2009. URL: http://www.nrel.gov/docs/fy10osti/47256.pdf.

[42] K. Deng et al. "Building thermal model reduction via aggregation of state". In: *American Control Conference*. 2010, pp. 5118–5123.

[43] Paul Denholm, Easan Drury, and Robert Margolis. *The Solar Deployment System (SolarDS) Model: Documentation and Sample Results*. Tech. rep. NREL/TP-6A2-45832. National Renewable Energy Laboratory, 2009. URL: http://www.nrel.gov/docs/fy10osti/45832.pdf.

[44] Paul Denholm et al. *The Value of Energy Storage for Grid Applications*. Tech. rep. NREL/TP-6A20-58465. National Renewable Energy Laboratory, 2013. URL: http://www.nrel.gov/docs/fy13osti/58465.pdf.

[45] Der Spiegel. *Germany's Energy Poverty: How Electricity Became a Luxury Good*. http://www.spiegel.de/international/germany/high-costs-and-errors-of-german-transition-to-renewable-energy-a-920288.html.

[46] Andre Dozier. "Integrated Water and Power Modeling Framework for Renewable Energy Integration". MA thesis. Fort Collins, CO, U.S.A.: Colorado State University, 2012.

[47] M. Duffy et al. "TRNSYS - Features and Functionality for Building Simulation 2009 Conference". In: *IBPSA Conference*. 2009, pp. 1950–1954.

[48] *Dynamic Demand*. http://www.dynamicdemand.co.uk/.

[49] Electric Power Research Institute. *Electricity Energy Storage Technology Options*. 2010. URL: http://www.electricitystorage.org/images/uploads/static_content/technology/resources/ESA_TR_5_11_EPRIStorageReport_Rastler.pdf.

[50] Energinet DK. *Environmental Report for Danish Electricity and CHP*. http://energinet.dk/SiteCollectionDocuments/Engelske\%20dokumenter/Klimaogmiljo/Environmental\%20report\%20for\%20Danish\%20electricity\%20and\%20CHP\%20-\%20summary\%20of\%20the\%20status\%20year\%202012.pdf.

[51] European Energy Exchange AG. *EEX Transparency Platform*. http://www.transparency.eex.com/en/.

[52] Jason Flinn and M. Satyanarayanan. "PowerScope: A Tool for Profiling the Energy Usage of Mobile Applications". In: *Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications*. 1999.

[53] Matthias Fripp. "Optimal Investment in Wind and Solar Power in California". PhD thesis. Energy and Resources Group, University of California, Berkeley, 2008.

[54] Matthias Fripp. "Switch: A Planning Tool for Power Systems with Large Shares of Intermittent Renewable Energy". In: *Environmental Science and Technology* 46.11 (2012), pp. 6371–6378.

[55] P. Gill, W. Murray, and M. Saunders. "SNOPT: An SQP algorithm for large-scale constrained optimization". In: *SIAM Review* 47.1 (2005), pp. 99–131.

[56] V W Goldschmidt. "Heat Pumps: Basics, Types, and Performance Characteristics". In: *Annual Review of Energy* 9.1 (1984), pp. 447–472. DOI: 10.1146/annurev.eg.09.110184.002311.

[57] Gene Grimm et al. "Examples when nonlinear model predictive control is nonrobust". In: *Automatica* 40.10 (2004), pp. 1729–1738.

[58] W. Guo and M. Zhou. "Technologies toward Thermal Comfort-based and Energy-efficient HVAC Systems: A Review". In: *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*. 2009.

[59] He Hao et al. "Aggregate Flexibility of Thermostatically Controlled Loads". In: *IEEE Transactions on Power Systems* (2013).

[60] Elaine K. Hart, Eric D. Stoutenburg, and Mark Z. Jacobson. "The Potential of Intermittent Renewables to Meet Electric Power Demand: Current Methods and Emerging Analytical Techniques". In: *Proceedings of the IEEE* 100.2 (2012), pp. 322–334.

[61] G. Henze, C. Felsmann, and G. Knabe. "Evaluation of optimal control for active and passive building thermal storage". In: *International Journal of Thermal Sciences* 43.2 (2004), pp. 173–183.

[62] Honeywell. *Honeywell Engineering Manual of Automatic Control for Commercial Buildings*. Honeywell, 1998.

[63] G. Hoste, M. Dvorak, and M. Jacobson. *Matching Hourly and Peak Demand by Combining Different Renewable Energy Sources: A Case Study for California in 2020*. 2009.

[64] Tyler Hoyt et al. "Energy Savings from Extended Air Temperature Setpoints and Reductions in Room Air Mixing". In: *Proceedings of the International Conference on Environmental Ergonomics 2009*. Boston, MA, USA, 2009.

[65] ICF Consulting. *Electricity Demand in Ontario A Retrospective Analysis*. 2005.

[66] Independent Electricity System Operator – Ontario. *Hourly Generator Output and Capability*. http://www.ieso.ca/imoweb/marketdata/genEnergy.asp.

[67] X. Jaing et al. "Design and Implementation of a High-Fidelity AC Metering Network". In: *The 8th ACM/IEEE International Conference on Information Processing in Sensor Networks Track on Sensor Platforms, Tools, and Design Methods (IPSN/SPOTS 09)*. 2009.

[68] Xiaofan Jiang et al. "Design and Implementation of a High-Fidelity AC Metering Network". In: *The 8th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN¡D5¿09) Track on Sensor Platforms, Tools, and Design Methods (SPOTS '09)*. 2009.

[69] Xiaofan Jiang et al. "Experiences with a High-Fidelity Wireless Building Energy Auditing Network". In: *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems (SenSys)*. 2009.

[70] T.M. Keep et al. "Using load switches to control aggregated electricity demand for load following and regulation". In: *Power and Energy Society General Meeting, 2011 IEEE*. 2011.

[71] J. Daniel Khazzoom. "Economic Implications of Mandated Efficiency in Standards for Household Appliances". In: *The Energy Journal* 1.4 (1980), pp. 21–40.

[72] J.H. Kim and W.B. Powell. "Optimal Energy Commitments with Storage and Intermittent Supply". In: *Operations Research* 59.6 (2011), pp. 1347–1360.

[73] Stephan Koch, Johanna L. Mathieu, and Duncan S. Callaway. "Modeling and Control of Aggregated Heterogeneous Thermostatically Controlled Loads for Ancillary Services". In: *Proceedings of the 17th Power Systems Computation Conference (PSCC)*. 2011.

[74] Andrew Krioukov et al. *Design and Evaluation of an Energy Agile Computing Cluster*. Tech. rep. UCB/EECS-2012-13. EECS Department, University of California, Berkeley, 2012. URL: http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-13.html.

[75] Andrew Krioukov et al. "NapSAC: Design and Implementation of a Power-Proportional Web Cluster". In: *Proceedings of the First ACM SIGCOMM Workshop on Green Networking*. 2010.

[76] M. Kvasnica, P. Grieder, and M. Baotić. "Multi-Parametric Toolbox (MPT)". In: (2004). URL: http://control.ee.ethz.ch/~mpt/.

[77] LabJack Corporation. *UE9 Data Acquisition Module*. http://labjack.com/ue9.

[78] W. Langson et al. "Robust model predictive control using tubes". In: *Automatica* 40.1 (2004), pp. 125–133.

[79] Lawrence Berkeley National Laboratory. *Appliance Energy Use Data Repository*. http://minotaur.lbl.gov/aeud/.

[80] C. Liao and P. Barooah. "An integrated approach to occupancy modeling and esimation in commerical buildings". In: *American Control Conference*. 2010, pp. 3130–3135.

[81] D. Limon et al. "Robust tube-based MPC for tracking of constrained linear systems with additive disturbances". In: *Journal of Process Control* 20.3 (2010), pp. 248–260.

[82] Zhenhua Liu et al. "Renewable and Cooling Aware Workload Management for Sustainable Data Centers". In: *Proceedings of ACM Sigmetrics*. 2012.

[83] Amory Lovins. "Energy Saving from the Adoption of More Efficient Appliances: Another View". In: *The Energy Journal* 9.2 (1988), pp. 155–170.

[84] Xi Lu, Michael B. McElroy, and Juha Kiviluoma. "Global Potential for Wind-generated Electricity". In: *PNAS* (2009). URL: http://www.pnas.org/content/early/2009/06/19/0904101106.full.pdf.

[85] T. J. Lui, W. Stirling, and H. O. Marcy. "Get Smart". In: *Power and Energy Magazine, IEEE* 8.3 (2010), pp. 66–78.

[86] John Lygeros, Claire Tomlin, and Shankar Sastry. "Controllers for Reachability Specifications for Hybrid Systems". In: *Automatica* 35.3 (Mar. 1999), pp. 349–370.

[87] Y. Ma, G. Anderson, and F. Borrelli. "A Distributed Predictive Control Approach to Building Temperature Regulation". In: *American Control Conference*. 2011.

[88] David MacKay. *Sustainable Energy – Without the Hot Air*. http://www.withouthotair.com, 2009.

[89] MIT Energy Initiative. *The Future of the Electric Grid*. http://web.mit.edu/mitei/research/studies/documents/electric-grid-2011/Electric_Grid_Full_Report.pdf. 2011.

[90] Nathan Murthy et al. "Energy-Agile Laptops: Demand Response of Mobile Plug Loads Using Sensor/Actuator Networks". In: *Third IEEE Int'l Conference on Smart Grid Communications (SmartGridComm'12)*. 2012.

[91] National Public Radio: StateImpact. *ERCOT: So, About that Chance of Rolling Blackouts*. http://stateimpact.npr.org/texas/2012/07/10/ercot-so-about-that-chance-of-rolling-blackouts/.

[92] National Renewable Energy Laboratory. *Levelized Cost of Energy Calculator*. http://www.nrel.gov/analysis/tech_lcoe.html.

[93] National Renewable Energy Laboratory. *Renewable Electricity Futures Study*. http://www.nrel.gov/analysis/re_futures/.

[94] Navigant Research. *Executive Summary: Electric Vehicle Market Forecasts – Global Forecasts for Light Duty Hybrid, Plug-in Hybrid, and Battery Electric Vehicles: 2013-2020*. http://www.navigantresearch.com/wp-assets/uploads/2013/06/EVMF-13-Executive-Summary.pdf.

[95] T. Nghiem and G. Pappas. "Receeding-horizon Supervisory control of Green Buildings". In: *American Control Conference*. 2011.

[96] F. Oldewurtel et al. "Energy Efficient Building Climate Control using Stochastic Model Predictive Control and Weather Predictions". In: *American Control Conference*. 2010, pp. 5100–5105.

[97] Ontario Ministry of Energy. *Ontario's Long-Term Energy Plan*. http://www.energy.gov.on.ca/docs/en/MEI_LTEP_en.pdf.

[98] Pacific Gas & Electric. *Electric Rate Schedules*. http://www.pge.com/tariffs/ERS.SHTML.

[99] Anthony Papavasiliou and Shmuel S. Oren. "Coupling Wind Generators with Deferrable Loads". In: *IEEE Energy 2030 Proceedings*. 2008.

[100] Phys.org. *German Greenhouse Gas Emissions Rose in 2012*. http://phys.org/news/2013-02-german-greenhouse-gas-emissions-rose.html.

[101] Joseph Polastre, Robert Szewczyk, and David Culler. "Telos: Enabling Ultra-Low Power Wireless Research". In: *The 4th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN'05) Track on Sensor Platforms, Tools, and Design Methods (SPOTS '05)*. 2005.

[102] PVOutput.org. *Canada Statistics*. http://pvoutput.org/map.jsp?country=44.

[103] R. Katz, D. Culler, S. Sanders, et al. "An Information-Centric Energy Infrastructure: The Berkeley View". In: *Sustainable Computing: Informatics and Systems* 1 (2011), pp. 7–22.

[104] Richard K. Miller & Associates. *Utility Energy Storage – Pumped Hydroelectric Storage*. http://www.rkma.com/utilityenergystorageSAMPLE.pdf.

[105] B. Roberts. "Capturing Grid Power". In: *Power and Energy Magazine, IEEE* 7.4 (2009), pp. 32–41.

[106] P. Robinson. "Root-N-Consistent Semiparametric Regression". In: *Econometrica* 56.4 (1988), pp. 931–954.

[107] M. Roozbehani, M.A. Dahleh, and S.K. Mitter. "Volatility of Power Grids Under Real-Time Pricing". In: *IEEE Transactions on Power Systems* 27.4 (2012), pp. 1926–1940.

[108] S. Shankar Sastry and Marc Bodson. *Adaptive Control: Stability, Convergence, and Robustness*. Prentice-Hall, 1989.

[109] Walter Short et al. *Regional Energy Deployment System (ReEDS)*. Tech. rep. NREL/TP-6A20-46534. National Renewable Energy Laboratory, 2011. URL: http://www.nrel.gov/analysis/reeds/pdfs/reeds_documentation.pdf.

[110] Solar Schoolhouse. *Rooftop Solar Traces*. http://view2.fatspaniel.net/SSH/MainView.jsp.

[111] *Standards Interpretation - 29 CFR, Part 1910.1000*. OSHA.

[112] *State of California Executive Order S-21-09*. http://gov38.ca.gov/index.php?/executive-order/13269/.

[113] T. Peffer, *et al.* "Deep Demand Response: The Case Study of the CITRIS Building at the University of California-Berkeley". In: *Proceedings of the ACEEE 2012 Summer Study on Energy Efficiency in Buildings*. 2012.

[114] Jay Taneja, David Culler, and Prabal Dutta. "Towards Cooperative Grids: Sensor/Actuator Networks for Renewables Integration". In: *Proceedings of the 1st IEEE Int'l Conference on Smart Grid Communications*. 2010.

[115] Jay Taneja, Randy Katz, and David Culler. "Defining CPS Challenges in a Sustainable Electricity Grid". In: *Proceedings of the Third ACM/IEEE International Conference on Cyber-Physical Systems*. 2012.

[116] Jay Taneja, Ken Lutz, and David Culler. "The Impact of Flexible Loads in Increasingly Renewable Grids". In: *Proceedings of the Fourth IEEE Int'l Conference on Smart Grid Communications (SmartGridComm'13)*. 2013.

[117] Jay Taneja et al. "A Comparative Study of High Renewables Penetration Electricity Grids". In: *Proceedings of the Fourth IEEE Int'l Conference on Smart Grid Communications (SmartGridComm'13)*. 2013.

[118] J. A. Taylor, D. S. Callaway, and K. Poolla. "Inventory Control of Storage in Distribution Systems". In: *American Control Conference*. 2012.

[119] Russ Tedrake. "LQR-Trees: Feedback motion planning on sparse randomized trees". In: *Robotics: Science and Systems*. 2009, pp. 17–24.

[120] Claire Tomlin, John Lygeros, and Shankar Sastry. "A Game Theoretic Approach to Controller Design for Hybrid Systems". In: *Proceedings of the IEEE* 88.7 (July 2000), pp. 949–969.

[121] U.S. Department of Energy - Office of Energy Efficiency and Renewable Energy. *EnergyPlus Energy Simulation Software*. http://apps1.eere.energy.gov/buildings/energyplus/energyplus_documentation.cfm.

[122] U.S. Energy Information Administration. *2014 Annual Energy Outlook*. http://www.eia.gov/forecasts/aeo/er/index.cfm.

[123] U.S. Energy Information Administration. *Residential Energy Consumption Survey*. http://www.eia.gov/consumption/residential/data/2009/. 2009.

[124] J. V. C. Vargas and J. A. R. Parise. "Simulation in transient regime of a heat pump with closed-loop and on-off control". In: *International Journal of Refrigeration* 18.4 (1995), pp. 235–243.

[125] William Stanley Jevons. *The Coal Question*. 2nd ed. Macmillan and Co., 1866.

[126] World Nuclear Association. *Nuclear Power in Germany*. http://www.world-nuclear.org/info/Country-Profiles/Countries-G-N/Germany/.

[127] L. Xie et al. "Wind Energy Integration in Power Systems: Operational Challenges and Possible Solutions". In: *Proceedings of IEEE: Special Issue on Network Systems Engineering for Meeting the Energy and Environment Dream* 99.1 (2011), pp. 214–232.

[128] J.X. Xu and Y. Tan. *Linear and nonlinear iterative learning control*. Springer, 2003.

[129] Rongxin Yin et al. "Scenario Analysis of Peak Demand Savings for Commercial Buildings with Thermal Mass in California". In: *Proceedings of the ACEEE 2010 Summer Study on Energy Efficiency in Buildings*. 2010.

[130] Lide Zhang et al. "Accurate Online Power Estimation and Automatic Battery Behavior Based Power Model Generation for Smartphones". In: *Proceedings of the Eighth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*. 2010.