Powering the Information Age: Metrics, Social Cost Optimization Strategies, and Indirect Effects Related to Data Center Energy Use

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Engineering and Public Policy

Nathaniel Charles Horner

M.E., Systems Engineering, University of Virginia
M.S., Computer Science, Johns Hopkins University
B.S., Computer Science, North Carolina State University
B.A., English, North Carolina State University

Carnegie Mellon University
Pittsburgh, PA

August 2016

Acknowledgments

I am extremely grateful to the members of my thesis committee, who have helped immensely with the completion of this work. Inês Azevedo has offered encouragement and guidance from the day I arrived in EPP, while patiently allowing me the freedom to find my own research path. Doug Sicker, despite being busy with the monumental task of assuming the EPP helm in 2014, quickly volunteered to join my committee and has provided a much-needed telecom perspective on this work. Scott Matthews provided firm grounding—some might say indoctrination—in the “EPP way” of solving complex socio-technical policy problems, and I am privileged to have taken the core EPP 702 course during his final year of teaching it. Yuvraj Agarwal from the School of Computer Science helped me formulate a topic that enabled me to link my computer science background with my interest in energy use. I have relied on Arman Shehabi of Lawrence Berkeley National Laboratory (LBNL) and Jonathan Koomey of Stanford University to provide an industry-informed perspective, and their thoughtful, pointed feedback on my work has made me a better researcher.

Several people from other institutions have helped by providing data and insight. Dale Sartor and Rod Mahdavi of LBNL provided data and useful discussions in support of Chapter 2. Bruce Maggs at Akamai Technologies and Balakrishnan Chandrasekaran at Duke University provided the web traffic data without which Chapter 4 would be much less interesting. EPP alumnus Kyle Siler-Evans helped me use and extend his earlier work on estimating marginal damage factors for electricity generation. I am immensely appreciative of a number of other individuals from government, industry, and academia for taking the time to discuss data center energy use with a graduate student making his initial forays into the topic.

Other faculty who have made my time in EPP an intellectual thrill include Alex Davis, whose particular brand of analytical rigor is an examplar to any serious analyst, and Jay Apt, whose uncanny ability to separate the wheat from the chaff has on more than one occasion helped me identify the crucial elements of a problem. Eden Fisher made my stint as a teaching assistant a true joy. Debbie Stine has offered a

useful perspective on how analytically elegant solutions will fare against the harsh policymaking realities of the beltway—a grounding from which I think all of us in academia can benefit. I am grateful to David Hounshell for shepherding me through my first paper [1] and to Francisco Veloso for bringing me to EPP before heading off to new challenges.

Any department would be hopelessly lost without a support staff to help graduate students navigate the bureaucracy of scheduling, expense reimbursements, procuring resources, tracking down professors, and generally adhering to institutional requirements. Vicki Finney, Steve Gradeck, Adam Loucks, Patti Steranchak, Barbara Bugosh, Jaime Montgomery, and Elisabeth Udyawar have always met my questions and needs with smiles and competent aid.

While Carnegie Mellon has been the latest stop on my academic journey, my success here has been built upon important formative experiences at other institutions. I am indebted to Joe Herkert, the former director of the Benjamin Franklin Scholars program at N.C. State University, for early training on the ethical and interdisciplinary context of the engineering profession and for pointing me towards EPP when I was first considering graduate programs. At the University of Virginia, Bill Scherer and Mike Smith have imbued the Accelerated Master's Program (AMP) in Systems Engineering with a uniquely rigorous intellectual character that made my year of weekends in Charlottesville one of the most challenging but also most fulfilling of my career. AMP provided me a solid basis in key analytical tools that made my transition to full-time graduate school much easier.

EPP students define the program's social identity and have created a culture that is, I think, unique within Carnegie Mellon. I have truly enjoyed our time together in the collegial environment of Hamburg Hall (constant state of renovation notwithstanding) and the world beyond. Parth Vaishnav has become a close colleague and friend through three USAEE case competitions (with such success that we were asked to take over writing the cases) and collaboration on several enjoyable projects. I have benefitted from his patient ear when I needed to vent, his level-headed perspective when I needed advice, and his knowledge of local cuisine when I needed lunch. Additionally, I am grateful to Casey, Frauke, Roger, Chin Yen, Alan, Eric, D.L., Will, and Olga for their friendship and support.

Finally, I am blessed with a wonderful family. My parents, Charles and Virginia, fostered in me an intellectual curiosity from an early age and tirelessly supported many academic and extracurricular endeavors. My brother, Robert, has been a role model, confidant, and the best friend anyone can wish for. And to my wife, Lauren: for allowing me to drag you north to the Rust Belt, for joining me in subsequently falling in love with this great city of Pittsburgh, for putting up with the schedule and occasional stress of graduate student life, and for your love and support during this adventure, I cannot thank you enough.

My work in EPP has been funded by the Department of Engineering and Public Policy and additionally supported by the Center for Climate and Energy Decision Making and the CMU Graduate Student Association. In particular, I am grateful to Granger Morgan and Doug Sicker for their generosity in supporting my studies in EPP.

Abstract

This dissertation contains three studies examining aspects of energy use by data centers and other information and communication technology (ICT) infrastructure necessary to support the electronic services that now form such a pervasive aspect of daily life. The energy consumption of ICT in general and data centers in particular has been of growing interest to both industry and the public, with continued calls for increased efficiency and greater focus on environmental impacts.

The first study examines the metrics used to assess data center energy performance and finds that power usage effectiveness (PUE), the *de facto* industry standard, only accounts for one of four critical aspects of data center energy performance. PUE measures the overhead of the facility infrastructure but does not consider the efficiency of the IT equipment, its utilization, or the emissions profile of the power source. As a result, PUE corresponds poorly with energy and carbon efficiency, as demonstrated using a small set of empirical data center energy use measurements.

The second study lays out a taxonomy of indirect energy impacts to help assess whether ICT's direct energy consumption is offset by its energy benefits, and concludes that ICT likely has a large *potential* net energy benefit, but that there is no consensus on the sign or magnitude of *actual* savings, which are largely dependent upon implementation details.

The third study estimates the potential of dynamic load shifting in a content distribution network to reduce both private costs and emissions-related externalities associated with electricity consumption. Utilizing variable marginal retail prices based on wholesale electricity markets and marginal damages estimated from emissions data in a cost-minimization model, the analysis finds that load shifting can either reduce data center power bills by approximately 25%–33% or avoid 30%–40% of public damages, while a range of joint cost minimization strategies enables simultaneous reduction of both private and public costs. The vast majority of these savings can be achieved even under existing bandwidth and network distance constraints, although current industry trends towards virtualization, energy efficiency, and green power may make load shifting less appealing.

Keywords: information and communication technology, energy consumption, green computing, data centers, optimization, indirect energy effects

Contents

List of Tables

List of Figures

List of Abbreviations

AC	alternating current
ADF	average damage factor
AEF	average emissions factor
AP2	Air Pollution Emissions Experiments and Policy Analysis model, version 2
API	application programming interface
APN	aggregate pricing node
B2B	business-to-business
B2C	business-to-customer
BA	balancing authority (also balancing area)
BEMS	building energy management system
CD	compact disc
CDF	cumulative distribution function
CDN	content distribution network (also content delivery network)
CEMS	Continuous Emissions Monitoring System
$\mathbf{CO_2}$	carbon dioxide
CUE	carbon usage effectiveness
DC	direct current
DCeP	data center energy productivity
DOE	Department of Energy
DAM	day-ahead market
DR	demand response
DSM	demand-side management
DVD	digital video disc
EASIUR	Estimating Air Pollution Social Impact Using Regression

ECOR	equivalence class of regions
eGRID	Emissions & Generation Resource Integrated Database
EIA	Energy Information Administration
EPA	Environmental Protection Agency
FERC	Federal Energy Regulatory Commission
GEC	green energy coefficient
GHG	greenhouse gas
HPC	high-performance computing
HVAC	heating, ventilating, and air conditioning
ICT	information and communication technology (also IT)
ISO	independent system operator
ISO–NE	Independent System Operator–New England
ISP	internet service provider
IT	information technology (also ICT)
ITUE	information technology usage effectiveness (also SPUE)
LBNL	Lawrence Berkeley National Laboratory
LCA	lifecycle analysis
LCOE	levelized cost of electricity
LMP	locational marginal price
LP	linear program
MDF	marginal damage factor
MEF	marginal emissions factor
MIP	mixed-integer program
MISO	Midcontinent Independent System Operator
MWh	megawatt-hour
$\mathbf{NO_x}$	nitrogen oxides
NERC	North American Electric Reliability Corporation
NYISO	New York Independent System Operator
PDU	power distribution unit
PJM	Pennsylvania–New Jersey–Maryland Interconnection

PM	particulate matter (subscript indicates particle size in micrometers)
POV	personally-owned vehicle
PSU	power supply unit
PUE	power usage effectiveness
RTM	real-time market
RTO	regional transmission organization
RTP	real-time pricing
RTT	round-trip time
SLA	service level agreement
SO_2	sulfur dioxide
SPP	Southwest Power Pool
SPUE	server power usage effectiveness (also ITUE)
TUE	total-power usage effectiveness
UPS	uninterruptable power supply
URL	uniform resource locator
VMT	vehicle miles traveled
WSC	warehouse-scale computer

Chapter 1

Introduction

The advent of the digital computer brought about the “Information Age,” an era in which information has come into its own as a valuable commodity. The accuracy, relevance, and timeliness of an organization’s information are—as they have always been—keys to its success. However, the higher speeds, greater traffic, and increased access on the “information superhighway” have made businesses hungrier for ever-increasing volumes of data.

Data services lie at the heart of operations for many companies and constitute a core product for others. Google indexes the web to provide search services to users while simultaneously collecting information about search activity to deliver advertising. Social networking entities such as Facebook collect personal data in exchange for hosting virtual communities aiding interaction among groups. Electronic retailers like Amazon.com house online inventories and bring buyers and sellers together. Data service providers such as Dropbox and Apple allow users to store their documents, files, and digital content “in the cloud,” a distributed storage network. Even traditional retailers like Walmart and Whole Foods use data-intensive processes to manage their inventories in real-time.

The hardware and software used to store, transmit, and utilize data to provide these e-services is collectively known as information technology (IT) or information and communication technology (ICT). ICT includes computers and software, mobile devices, and communication networks and their components. As digital content has proliferated, so too have the storage mechanisms grown, moving from the lone server to the server closet, the server room, and now the server farm. These storage repositories are collectively known as data centers, which not only provide static storage but also dynamically provide a wide variety of services including hosting web pages and email, streaming multimedia, and running complex business applications like banking management software. Individuals, private firms, universities, and government entities all use data centers of varying scale and complexity to manage the digital information they need to operate.

1.1 Scope of the dissertation

The energy consumption of ICT in general and data centers in particular is of increasing interest to both industry and the public. The title of this thesis refers to three important aspects of data center energy use: how we measure it, how consuming energy in a data center might (or might not) enable energy savings in other sectors of society and the economy, and how we might reduce the private and social costs associated with this energy consumption.

Chapters 2 and 3 introduce and discuss the direct and indirect energy use of data centers and other ICT infrastructure. Chapter 2 focuses on the data center itself, reviewing metrics for assessing data center efficiency and using an empirical data set to illustrate why a current prevalent industry standard is not comprehensive enough.

Chapter 3 reviews current thinking on the net energy impacts of ICT as a whole; that is, does the proliferation of ICT yield indirect energy reductions in other areas of society and the economy, and, if so, do those reductions offset the energy consumed directly by the ICT equipment?

Chapter 4, which contains the main analytical content of the dissertation, evaluates an opportunity to reduce data center energy costs by shifting the computing load around a network of data centers. Importantly, this analysis considers both the private costs (i.e., the data center operator's power bill) and the external costs of energy (the impact on the public associated with electricity generation) and addresses whether these different goals lead to different load shifting strategies.

Chapter 5 synthesizes these studies and focuses on resultant policy recommendations. Before launching into these three studies, however, a brief introduction to data centers and their energy consumption is in order.

1.2 What is a data center?

The term *data center* encompasses a fairly heterogeneous group of facilities. Lawrence Berkeley National Laboratory (LBNL) describes a data center as a special-purpose facility with the following characteristics and functions [2]:

- “Houses various equipment, such as computers, servers (e.g., web servers, application servers, database servers), switches, routers, data storage devices, load balancers, wire cages or closets, vaults, racks, and related equipment.
- Store[s], manage[s], processe[s], and exchange[s] digital data and information;
- Provide[s] application services or management for various data processing, such as web hosting internet, intranet, telecommunication, and information technology.”

While the LBNL definition is focused on facility contents and function, ICT consulting firm Gartner uses a definition more focused on the organizational role the data center plays, defining a data center as:

> “the department of an enterprise that houses and maintains the back-end [IT] systems and data stores—its mainframes, servers, and databases. In the days of large, centralized IT operations, this department and all the systems resided in one physical space.” [3]

These definitions make no statements about size, and include few restrictions on function. Data centers can range from small server closets to huge server farms and can host a variety of IT services, such as corporate email and filesystems, data archives, and cloud services. The main criterion for a data center seems to be that it houses “back-end” ICT equipment—equipment accessed indirectly by users via a network. This variety of size and function can make clear classification difficult, though most taxonomies rely on some combination of size, criticality of service, and service type.

The Uptime Institute, an industry stakeholder group, focuses its data center classification on infrastructure redundancy, ranging from “basic” to “fault-tolerant” [4, 5]. Tier I data centers have no redundant systems, whereas Tier IV facilities have duplicate active power and cooling distribution paths, with redundant components on each, so that the center can withstand any single equipment failure.

We adopt the a slightly broader hierarchy, loosely based on the taxonomy used by IDC [6]:

1. *Server closets*, or “ad-hoc” data centers, support small businesses or individual projects at larger companies. They may get some support from a corporate-level IT department but may also be configured and operated by non-experts.
2. *Server rooms* are small data centers that support small businesses or special groups or projects of larger entities. They may be administered by central IT staff or “owned“ by each project or division.
3. *Localized data centers* provide business-critical applications and have some power and cooling redundancy, though downtime is not catastrophic. Restoration of service on the order of hours is acceptable.

4. *Mid-tier data centers* are medium-to-large data centers used to host enterprise-wide applications in support of operations or human resources (e-mail accounts, filesystems, internal data). The data is critical, but incidental to the primary business line. Downtime lasting longer than a few minutes has significant impact on the business. These facilities are operated by the company's central IT department.
5. *Enterprise data centers* are large facilities used, usually by non-ITC companies, in support of core business operations (e.g., banks, health care companies, etc.). These data centers are often in special-purpose facilities and operated under a separate business unit or division. Downtime is catastrophic, and these facilities have highly redundant infrastructure.
6. *Hyperscale data centers, server farms, or warehouse scale computers (WSCs)* are the very large data centers, usually constructed in their own physical plants, built by ICT companies with a primary business line focused on data (e.g., Google, Apple, Facebook, Amazon, etc.) and, increasingly, cloud-based services.Barroso and Hölzle [7] coined the term Warehouse Scale Computer to emphasize the distinguishing large economies of scale, extreme parallelism, hardware and software homogeneity, and aggressive focus on efficiency of these data centers.

Generally, size, infrastructure redundancy, quality of service, and criticality increase as one moves down the list, though these distinctions are necessarily qualitative and somewhat fuzzy in nature. For an even more granular data center typology, see Table 2 in Shehabi *et al.* [8]. These data center types can host different types of applications in very different domains, ranging from corporate entities to university and research-based enterprises.

1.3 Energy consumption by data centers

While there are several different options for configuring a data center's energy path, electricity enters a "typical" data center as alternating current (AC) from the grid (or a backup generator) and then flows through the following equipment: a bank of backup batteries called uninterruptable power supplies (UPSs), where it is converted to direct current (DC) and back to AC; power distribution units (PDUs) to route power to different loads; transformers to step down the voltage; power supply units (PSUs), which convert AC to DC; and finally to the servers themselves, where energy is consumed by processors (CPUs), memory, disks, and other hardware components. Energy is also used to run networking equipment, lighting, and cooling systems. Cooling can consume as much energy as the servers in an inefficient data center, and design concepts such as hot-aisle/cold-aisle separation between server banks reduce the proportion of total energy used for air conditioning. The UPSs provide short-term backup power, while longer-term backup is often met with diesel generators. (Chapter 4 in Barroso and Hölzle [7] provides a more comprehensive overview.)

The possible variations on this typical setup are myriad and include using distributed, in-rack UPSs [9], taking advantage of "free cooling" by using ambient air where environmental conditions allow [10], converting to DC power early in the power path to reduce losses [11–13], and co-locating distributed renewable generation at the data center facility [14]. We will discuss how some of these options affect data center energy use in Chapter 2.

Expansion of the data used and services provided by these enterprises has meant exponential growth in needed storage capacity. IBM estimated global daily data production at 2.5 quintillion bytes per day in 2013 and reported that 90% of the world's data had been produced in the previous two years [15], and of course these figures are already dated. In addition to the magnitude of data produced, increasing complexity of software has increased its size. As the prevalence of data centers has grown, so have public concerns about their aggregate energy consumption [16, 17], though objective assessment of some of the more sensational claims on this topic shows them to be egregiously overblown [18–20].

A series of bottom-up estimates[1] has generally found that data centers use on the order of 1-2% of U.S. and global electricity consumption. Data center energy use in the U.S. nearly doubled between 2000 and 2005, but this growth slowed from 2005-2010, increasing by just over a third during that time, and has further flattened since then, increasing only 4% from 2010-2014 [8, 22, 23].

Despite increasing efficiency and slowing energy consumption growth in data centers, the decisions governing the ways in which ICT is deployed and operated will dictate whether this new infrastructure becomes an energy hog or an important tool in achieving energy reduction goals. This dissertation highlights three areas of different scope—the individual data center, the network, and the broader ICT service infrastructure—where these critical decisions are being made.

[1] These estimates are "bottom-up" in the sense that they estimate the power consumption of individual IT components (e.g., server types), factoring in utilization and deployment characteristics such as PUE (see Chapter 2), and sum over the estimated installed base to get a value for overall annual energy consumption. This estimate can be used as an input for ICT energy intensity metrics using what Aslan *et al.* [21] call the *annual electricity consumption* method.

Chapter 2

Data Center Efficiency Metrics

Some of the content in this and the preceding chapter has been published in N. Horner and I. Azevedo, "Power usage effectiveness in data centers: Overloaded and underachieving," The Electricity Journal*, vol. 29, no. 4, pp. 61–69, May 2016. [24]*

> **Motivating questions: what are the current metrics used for assessing data center energy performance? Is there evidence that current metrics are not adequately aligned to energy efficiency and "green" computing?**

Before undertaking involved—and potentially costly—interventions to reduce energy use in data centers, it is clearly important to understand the nature and context of the problem such interventions hope to solve. In other words, are data centers actually inefficient in their use of electricity? If so, in which components of the data center system do the inefficiencies reside?

ICT industry groups and government-affiliated research centers have worked to establish metrics to answer these questions, to respond to energy and environmental concerns, and to identify potential operating cost savings. This chapter provides a general overview of the metrics used to assess data centers' energy performance, including a critique of the most prevalent metric, power usage effectiveness (PUE).

2.1 The low-carbon ideal: energy productivity

Before examining existing metrics, we will consider what the "ideal" data center efficiency metric would look like. The best energy efficiency metric should measure the amount of useful computational work performed per unit of energy used; i.e., it would be a *productivity* metric:

$$\eta_E = \frac{\text{useful computational work}}{\text{energy consumed}} \tag{2.1}$$

If we are interested in carbon efficiency or another measure related to an externality, we could replace the denominator with *carbon emitted* or another measure of energy impact.

Indeed, the Green Grid has proposed a Data Center Energy Productivity (DCeP) metric along these lines [25]. The challenge, of course, is in defining "useful work" in a meaningful, measurable way.

EBay's Digital Service Efficiency dashboard [26] (which posted current data in 2012 and 2013, but no longer appears to be updated) includes such productivity metrics, reporting URLs per kWh and revenue per server, per user, and per MWh. The dashboard also reports traditional metrics, such as PUE (which we will discuss in detail below). Work commissioned by Salesforce.com evaluates carbon efficiency on a per-transaction basis, where transactions are defined as either web or application programming interface (API) requests—although even this definition is an abstraction, "given the diversity and complexity in types of transactions, but lack of a standardized methodology to separate [them by] type" [27]. Other industry-specific metrics, such as energy per web search, per bank transaction, per e-mail, per sale, or per user account, could be envisioned. The drawback of such special purpose metrics is that they do not support comparison of data centers that serve different types of loads.

2.2 Energy productivity as a composition of other metrics

Because of the difficulty in defining data center energy productivity in widely-applicable terms, the approach taken by industry has been to develop a set of intermediate metrics that, when composed, align with data center productivity, energy-efficiency, and carbon efficiency.

Masanet *et al.* [28] lay out a conceptual energy-carbon performance map, identifying the efficiency of the data center facility, the efficiency of the IT equipment, and the carbon emissions profile of the power source as the three major determinants of data center carbon performance. A white paper from the Uptime Institute defines four factors in data center efficiency [29], each of which represents a different level of energy "overhead" and can be measured with different metrics (Table 2.1). The three components of low-carbon data centers identified in Masanet *et al.* [28] map to the facility, IT, and strategic levels in the Uptime Institute report.

Table 2.1: Components of data center energy productivity. Columns 1 and 2 are defined in Stanley *et al.* [29]. Column 3 relates each factor to the scope of efficiency or level of overhead it addresses, while Column 4 maps various metrics, defined below, to each factor.

Factor	**Description**	**Efficiency scope**	**Example metrics**
Physical site infrastructure overhead	Facility siting, design, construction, and operation	Facility	PUE
IT hardware energy efficiency	Efficiency of servers and hardware components (drives, chips, power supplies, etc.)	IT	SPUE, ITUE
IT hardware asset utilization	Use of available capacity	Operational	Utilization
IT strategy	Integration and alignment of data center operations with business goals	Strategic	GEC

Barroso and Hölzle [7] factor DCeP into three individual components:

$$\begin{aligned}\text{DCeP} &= \frac{1}{\text{PUE}} * \frac{1}{\text{SPUE}} * \frac{\text{computation}}{\text{energy to electronic components}} &(2.2)\\ &= \frac{1}{\text{TPUE}} * \frac{\text{computation}}{\text{energy to electronic components}} &(2.3)\end{aligned}$$

The three terms in this equation measure energy overhead at different levels in the data center, in order of increasing closeness to the computational hardware from left to right. The first term accounts for facility efficiency as PUE. PUE is the ratio of total power used by the data center facility to the power used by the IT equipment:

$$\text{PUE} = \frac{\text{total facility power}}{\text{IT equipment power}} \tag{2.4}$$

Thus, lower PUE values are better, with 1.0 being ideal. A PUE of 1.0 would indicate that 100% of the power delivered to the facility is used by the computing equipment. Power used for lighting, cooling, and other overhead increases PUE. Some have argued that 1.0 is not necessarily a minimum value, as use of recovered waste heat could enable PUE ratings of less than 1.0 [30]—although this is perhaps stretching the definition of PUE to encompass energy performance issues beyond its intended application.

The second term in Equation 2.3 addresses part of IT hardware efficiency using server power usage effectiveness (SPUE), also called information technology usage effectiveness (ITUE), which is analogous to PUE but focuses on how power is used at the server level—i.e., whether it is used by the computing components (e.g., CPUs and memory) or supporting infrastructure (e.g., power supplies and cooling fans). In short, PUE is a measure of *facility* overhead, whereas SPUE is a measure of *IT equipment* overhead. Different data center architectures blur the line between facility infrastructure and IT infrastructure by aggregating cooling fans and power supplies outside of each server box or, going the other way, distributing power backup batteries at the server level. Since it is not always clear on which side of the "IT boundary" components reside, total-power usage effectiveness (TUE), as the product of PUE and SPUE, captures both jointly [31].

The third term in Equation 2.3 addresses the efficiency of the internal server hardware itself: processors, memory, and other electronic components. Metering consumption at this level is impractical, although the energy efficiency of these components can be assessed in lab settings. Just as hardware can be benchmarked for performance, protocols such as JouleSort [32] benchmark energy use by running pre-determined workloads on the hardware while measuring energy consumption. With such benchmarks, we can get some sense of the "absolute" or "theoretical" energy efficiency of the data center as constructed and provisioned.

The components of the DCeP equation, representing facility and IT overhead, focus on what might be thought of as the physical plant of the data center: the aspects of energy use related to design, equipment selection, and how the power and cooling infrastructure is operated. However, the energy efficiency of a data center is also heavily affected by how the IT resources are used operationally. The third factor in Table 2.1 refers to what might be called *operational* overhead and has to do principally with server utilization. Because data center capacities are often sized for peak load to ensure high quality-of-service, most of the time servers are severely underutilized. Since operators are concerned about handling load spikes, substantial safety margins even beyond the observed peak load are common [33]. A measurement of 5,000 Google data centers revealed utilization of 60% or less 95% of the time and of 30% or less half of the time [7]. Importantly, Google represents the upper end of utilization; most data centers will see much lower utilization rates. The issue with underutilization is that servers are not energy proportional—that is, energy usage does not scale linearly with computing load. An x86 server at idle consumes almost 50% of its peak power usage [7, fig. 5.8], though more recent servers achieve better performance. Therefore, idle power consumption can be reduced in two ways: reducing the number of idle servers, and working towards energy-proportional hardware. We further discuss methods for increasing utilization in Section 2.4.3.

Finally, the fourth item in Table 2.1 is *strategic* overhead, which assesses whether the organizational decisions made from a systems perspective, such as procurement, IT system design, and operational and management processes, are supportive of energy-efficient data center operations. There are no standard metrics for strategic overhead, though analysis of industry leaders can reveal corporate best practices for data center management. The eBay case study by Schuetz *et al.* [34], which identified hardware standardization, high rack density, resiliency, redundancy, power efficiency, high utilization, and a focus on cost effectiveness as key aspects of data center infrastructure design—is one example of such an analysis.

All of these metrics by definition focus on energy efficiency, which is correlated with both "greenness" and cost savings. However, if the ultimate goal is to measure environmental footprint, energy efficiency is only half of the picture: reducing the carbon intensity of the power supply is also important. Some companies do focus on emissions, and there are "sustainablility metrics" designed to evaluate carbon performance, such as Carbon Usage Effectiveness (CUE), which is essentially PUE multiplied by a carbon emission factor [35]:

$$\text{CUE} = \frac{\text{CO}_2 \text{ emissions from total facility energy consumption}}{\text{IT equipment energy}} \tag{2.5}$$

When it was active, eBay's dashboard [26] reported metric tons of carbon per MWh, per server, and per million active buyers. Other companies, such as Facebook and Apple, report carbon usage to some extent in annual environmental footprint statements. Data centers with the same PUE or DCeP could have vastly different CUEs. For instance, consider Apple's data center in Maiden, NC, which is co-located with the country's largest non-utility-owned solar farm and largest non-utility-owned fuel cell plant (Apple 2014). The very same data center would have a much different carbon performance if it were powered by grid electricity. Another proposed metric is the Green Energy Coefficient (GEC), which is the ratio of green energy to total energy consumed by the data center [36].

Primary power reliability also plays a role in a data center's operational carbon emissions, since interruptions in the main supply are generally met with backup diesel generation. In Northern Virginia, another data center hub, the aggregate capacity of diesel generators is nearly equivalent to that of a nuclear power plant [17]. Furthermore, evaluating the environmental footprint of a data center requires a lifecycle-cost analysis (LCA) approach to include the embodied emissions of the facility and IT equipment [37], although these carbon emissions are likely small compared to operational emissions [28].

Finally, while these metrics are among the most prevalent, many others exist. Jamalzadeh and Behravan [38] list no less than thirty metrics proposed by various organizations to measure different aspects of data center energy efficiency [38].

2.3 Overextending PUE

It should be clear that PUE, at best, only measures one of four critical components of data center energy performance (Table 2.1). However, reporting of PUE has become the *de facto* standard [39]—or perhaps more than simply *de facto*: the EPA used PUE as the basis for its ENERGY STAR for Data Centers rating program, started in 2009 [40]—and has led to a sort of arms race among large data center operators to report the lowest PUE value.

Companies like Google and Facebook have aggressively reduced their PUE in recent years through a focus on efficiency and custom hardware design. Google reports a trailing twelve-month (TTM) fleet average PUE of 1.12, with individual site ranges from 1.09 to 1.31 [41]. Facebook does not report a fleet average, but provides dashboards showing real-time PUE measurements for two of its largest data centers, which report TTM averages of 1.08 and 1.09 [42, 43].

In contrast, businesses where data centers are a more ancillary part of operations likely have higher PUEs, and small-to-medium data center operators are less likely to focus intensely on the energy efficiency of their installations [44]. According to an Uptime Institute survey, among firms with fewer than 1,000 servers, only 50% of operators were concerned with PUE; among firms with over 5,000 servers, that number was 90% [45]. The smaller firms do not have the resources or expertise to achieve the ultra-low PUE benchmarks set by the big firms, and the lower financial rewards put data center greening off the radar of executive leadership. Reported industry-wide average PUE measures, based on survey data, vary between 1.8 [46] and 2.9 [47]. The latter survey reported that only 20% of data centers had PUEs of less than 2.0. In these instances, assessing PUE can identify "low-hanging" fruit that can lead to important facility-level improvements in these underperforming data centers.

A main reason for PUE's prevalence is its relative ease of measurement—although even that is not as straightforward as it may first seem. However, PUE is overemphasized and overextended. The issue is not the use of PUE in and of itself; PUE is a reasonable and helpful metric for assessing the facility infrastructure efficiency of a data center. Rather, the issue is that PUE is far from a comprehensive metric, and quoting PUE values *and nothing else* stops short of providing a truly meaningful sense of data center energy and carbon performance.

While PUE is understood to be but one of several interrelated factors that determine energy and carbon performance by many in the industry, the metric continues to be emphasized. According to respondents to recent industry surveys by the Uptime Institute, over 80% of IT executives track and report PUE to corporate management [48], and it was ranked as the most important data center metric. Utilization ranked near the middle of the metrics list, while carbon emissions were dead last [49]. Further, many data center operators wrongly report PUE as a proxy for environmental performance, industry press has conflated low PUE values with greenness [50], and even the GHG Protocol draft reporting standard for data centers lists only PUE in its section on calculating operational emissions [51]. Some governments, notably Amsterdam, a data center hub, have enacted legislation setting maximum PUE standards in a push to be green [52, 53], while in the U.S., PUE has been adopted as the basis for the ENERGY STAR data center labeling program [40]. Thus, the problems resulting from such overreach bear repeating.

2.3.1 PUE as a measure of facility and equipment efficiency

Facility efficiency is exactly what PUE is designed to measure. PUE is improved by reducing overhead (e.g., cooling, lighting) in comparison to the compute load (IT equipment). Calculating PUE requires two measures of power use: total facility power in the numerator and IT equipment power in the denominator. While the calculation is a simple one, in practice there is variability resulting from differences in measurement points.

Total facility power is intended to be measured at the utility meter [39]. However, smaller data centers inside multiuse facilities may not be independently metered, while very large data centers may be metered at higher voltages. In the latter case, measuring at the meter will impact PUE by including losses associated with step-down transformers that occur outside the scope of measurement for other data centers. In the former case, it can be difficult to draw a clear boundary around the system. If cooling, lighting, and HVAC systems are shared between the data center and spaces dedicated to other uses (e.g., offices), which is especially common in the smallest data centers, then it can be hard to accurately assess energy performance. Inconsistency in what is included in the measurements can make comparisons among different data centers difficult.

The IT equipment power can be affected by what is included as “IT equipment.” Though they are viewed as infrastructure, components like cooling fans and power supply units may be either counted as IT equipment or not, depending on whether they are housed internally or externally to servers. To get a “correct” value for IT equipment energy, measurements would need to be taken at the component level:

CPU and other integrated circuits, memory, disks, et cetera [30]. The fact that such measurements are impractical means that equipment efficiency can muddy the PUE calculations. A low-overhead facility running older, less efficient servers could conceivably achieve a low PUE while still using more energy than it needs. In practice, this may not be much of a concern for larger data centers that refresh their server stock frequently. However, data centers may need to look beyond PUE to ensure that the hardware they use is matched to their computing needs and is energy-efficient.

The variation in both facility and IT equipment boundaries means that PUE measures may not be consistent or directly comparable and provides opportunities for organizations to game the rating. Furthermore, actions that improve energy efficiency can perversely *increase* PUE—reducing IT load through virtualization without a parallel reduction in infrastructure load, for instance [54]. Finally, PUE does not capture the efficiency of the IT equipment itself. (See Equation 2.3.)

2.3.2 PUE as a measure of business operational and strategic efficiency

It is worth noting that energy efficiency is not the most important measure of effectiveness for data centers, which necessarily prioritize access time, availability, or other such "quality-of-service" metrics. In the early rush to build out storage capacity, energy efficiency was not a primary concern for data center operators. A focus on performance, server uptime, and hardware costs by IT engineers who purchased the equipment left operating costs as an afterthought. Even at companies like Google, an aggressive leader in data center energy efficiency, "it was clear the only way to make [search] work as [a] free product was to run on relatively cheap hardware" according to Urs Hoelzle, the company's vice president of operations [qtd. in 55].

However, the growth of data centers means that their operating budgets are an increasing part of overall corporate spending. A heightened focus on efficiency has led to declining PUEs. Unfortunately, for companies focused on improving operational efficiency, PUE says nothing about how well energy is being translated into the services or products the organization delivers. A company operating a server farm with a very low PUE but without an optimal allocation of the computing load to the hardware may be operationally inefficient. Such inefficiencies can result from excessive redundancy in the system or underutilization of the hardware [56–58].

2.3.3 PUE as a measure of "greenness"

Another reason to care about data center energy use—and the one most likely to be at the forefront of interactions with the public—is in the context of "green" operations. Measures that increase energy efficiency to reduce operating costs also tend to reduce greenhouse gas (GHG) emissions, criteria air pollutants, and impacts on water. McKinsey & Co. [58] estimated that in 2007 data centers were responsible for 170 Mt CO_2 worldwide and projected emissions to quadruple by 2020. GHG emissions attributable to data centers come from four sources: (1) electricity consumption, (2) onsite combustion, (3) refrigerant use, and (4) embodied emissions [51]. We discuss PUE in the context of (1) and (2); (3) and (4) are important factors in overall data center environmental evaluation but are outside the scope of this work.

Greenpeace's 2012 report, "How clean is your cloud?" noted that GHG emissions associated with electricity supply are important, and claimed that several of the largest data center operators relied heavily on dirty electricity [59]. In addressing data center professionals at the Uptime Symposium, the author of the Greenpeace report noted that "PUE is a very useful metric and diagnostic tool, but it is not a good indicator of how green you are . . . . It does not speak to the resources you use in the outside world" [qtd. in 14].

We now illustrate that PUE does not necessarily correlate with carbon efficiency using a small sample of data center measurements. Most companies that measure their data center performance hold results as proprietary information. Here, we use data from 32 data centers in a series of LBNL benchmarking studies [e.g., 60, 61], supplemented with analysis of two federal data centers owned by the Environmental Protection Agency [6] and the National Renewable Energy Laboratory [62] as well as self-reported and estimated data for four WSC-type data centers operated by Apple (two) and Facebook (two).

The LBNL case studies were focused on cooling, but they also reported power consumption broken down by use: computing, HVAC, lighting, and UPS losses. We estimate PUE for each facility by dividing total energy use by energy used for computing. As these case studies were carried out by three different organizations, there is some variation in the data reported and the terminology used, particularly with respect to facility size as distinct from building size. Many, though not all, of the data centers occupy multiuse buildings. We made a best effort to calculate PUE based on only the data center portion of the facility (i.e., the "white space," or the space inside the data center cooling envelope used by the IT equipment).

For Facebook, self-reported PUE values were used for the Forest City, NC and Prineville, OR data centers. Apple does not publish PUE values, though one data center industry insider claims that Apple's Maiden, NC facility rates at about 1.1 [50]. Interestingly, this article implies that measures like the solar arrays and fuel cell plant at the facility give it "an advantage" in the "constant PUE chase." This false conflation of PUE values with carbon efficiency by a data center professional is troublesome but not all that uncommon. Greenpeace has estimated the PUE for the facility at 1.35 [63], though there is considerable controversy over its assumptions. We use the 1.1 number for Apple's Maiden facility under the assumption that the Greenpeace estimate is too pessimistic and that Apple is generally competitive with Facebook and Google. We use the 1.35 value for Apple's older Newark, CA data center. However, given the rapidly changing landscape and lack of detail in the data provided, it would be best to treat these as abstract warehouse-scale data centers rather than accurate representations of specific Apple and Facebook facilities.

Figure 2.1 shows PUE plotted against physical size for these data centers. Unsurprisingly, the large WSC data centers have lower PUE ratings. However, PUE is less correlated with size for other data center tiers. It is important to note that most of the LBNL case studies are now at least ten years old, so, although industry reports indicate that average PUE values may still be near 2.0 and that the rate of improvement has stagnated [47, 48], the worst performers may have subsequently adopted best practices and improved their PUE ratings.

One particular design innovation not reflected in the data centers in Figure 2.1 that helps drive smaller-scale data centers to lower PUEs is modularization or, specifically, containerization [64], which allows the energy overhead to scale more closely in parallel with the IT load. A controlled comparison of modular vs. traditional raised-floor data centers carried out by colocation provider IO found that the modularized design had 44% lower energy overhead, with a PUE of 1.4 compared to a PUE of 1.7 in the traditional facility [65].

Figure 2.2 shows whether PUE or size is correlated with data center energy or carbon performance. The horizontal axis is energy intensity (energy used per square foot); the vertical axis is carbon intensity. Bubble area corresponds to data center size in white space square footage, while bubble color reflects the PUE of the data center, with lower PUEs (light green) being better, and higher PUEs (dark blue) being worse. Carbon emissions factors, based on the NERC region in which each facility is located, are from the Greenhouse Gas Protocol [66]. The data set has a wide variation in data center size, type, and use.

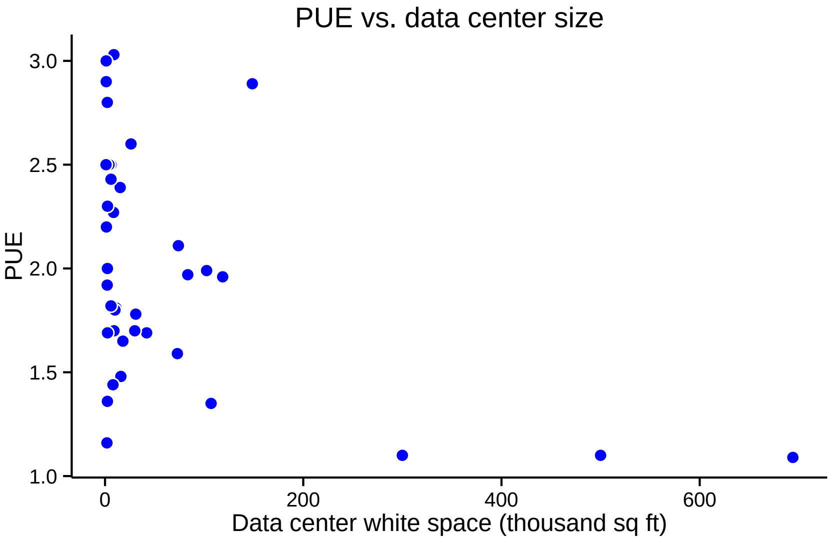


Figure 2.1: Data center PUE vs. physical size (white space floor area) for 38 heterogeneous data centers. Data are from LBNL bench-marking studies; one EPA legacy data center; one state-of-the-art data center built at NREL in 2010; and 2012 data from four WSC-type data centers.

Facilities are located in twelve states, with over-representation in California and Hawaii.

We note first that energy intensity is highly correlated with carbon intensity. This relationship makes sense, especially since GHG emissions are calculated directly from energy use. We expect greener data centers to lie along a "flatter" slope. However, PUE does not generally identify clusters of data centers by either measure of performance. While low-intensity data centers do tend to have poor PUE measures, three facilities at the opposite end of the spectrum also show relatively low carbon and energy intensity. The point is made particularly clear by comparing data centers along the 0.5 MWh/sq ft line; data centers with very different PUE ratings have the same energy intensity, while the smallest of these has much worse carbon performance despite having the better PUE value.

Figure 2.2 demonstrates empirically what Masanet *et al.* [28] discuss notionally. Their Figure 2 shows that two data centers with the same PUE can have very different energy use and carbon emissions profiles due to differences in power source and the efficiency of the IT equipment.

Precise metrics that go beyond PUE to address the full range of data center energy performance issues can highlight the aspects of the system where investment is most likely to yield energy, carbon, and cost savings. We now turn to a brief discussion of some of these target areas.

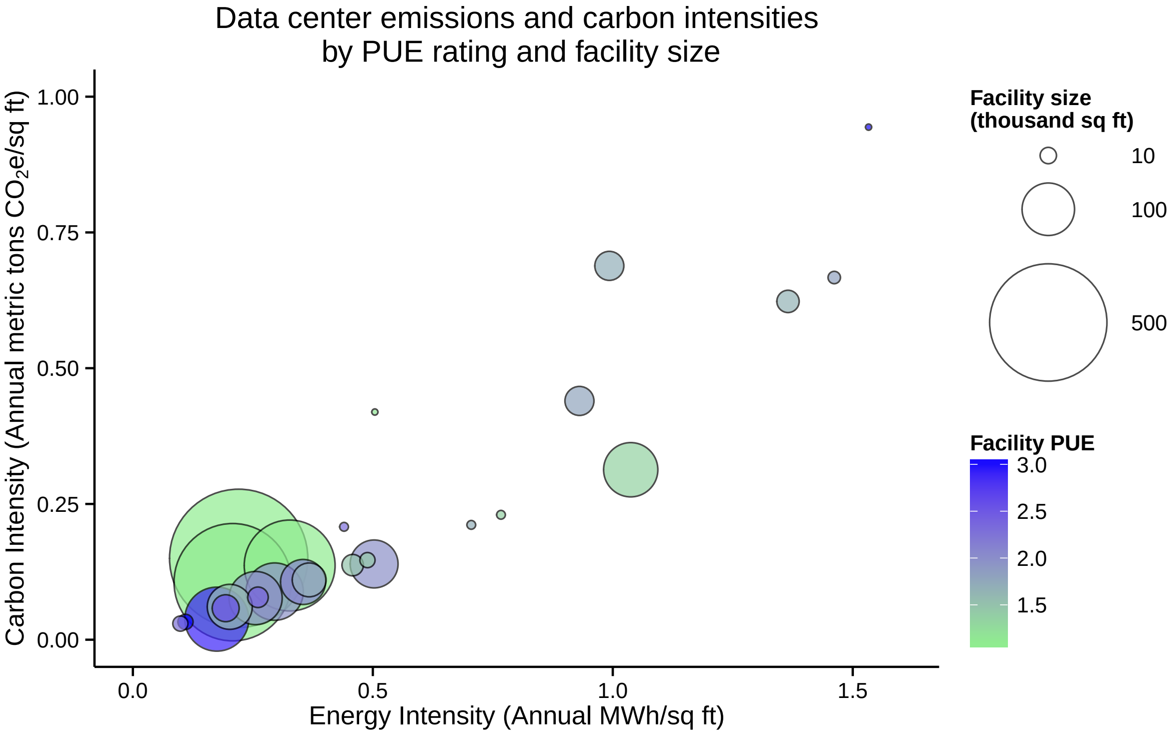


Figure 2.2: Annual energy intensity (in MWh/sq ft) vs. GHG emissions intensity (in metric tons CO_2e/sq feet) for 26 data centers in the United States. Bubble area corresponds to data center in white space size square footage; color scale identifies measured PUE. Included data centers are from LBNL benchmarking studies, one EPA legacy data center, one state-of-the-art data center built at NREL in 2010, and 2012 data from four WSC-type data centers. GHG for all data centers except for the WSC facilities are calculated from GHG Protocol grid emissions factors for the year closest to the energy usage report; GHG for the WSC facilities are self-reported by the data center owner.

2.4 Improving data center energy performance

There have been many suggested strategies for improving the energy and carbon performance of data centers, but they fit into the broad categories of *building efficiency*, *equipment efficiency*, *equipment utilization*, and *power sourcing*—which, not coincidentally, mostly align with the overhead categories in Table 2.1.

2.4.1 Building efficiency

Putting data centers in facilities specifically designed for them improves their efficiency. Standard practices include physical separation of the "hot" and "cold" aisles between server racks, raised floors, and carefully designed cabling conduits. An important contributor to efficiency is initial site selection: data centers in cooler climates or unique locations (e.g., underground or near water sources) allow for "free cooling." Some advanced data centers reduce the power loss by distributing DC current at the facility level, rather than converting AC to DC in each individual power supply unit.

2.4.2 Equipment efficiency

Most studies seem to identify other sources of efficiency gains as more important than addressing the efficiency of the IT hardware itself. However, the EPA ENERGY STAR program has issued product specifications for enterprise servers, uninterruptable power supplies (UPSs), data center storage, and network equipment [67]. Furthermore, if energy-proportional equipment could be achieved, utilization rates would no longer matter, since power usage would scale linearly with load.

2.4.3 Equipment utilization

Server utilization in data centers is generally very low. Because uptime, reliability, and fulfillment of service level agreements are the priorities of data center operators, data centers are generally built with extreme redundancy: "McKinsey & Company analyzed energy use by data centers and found that, on average, they were using only 6 percent to 12 percent of the electricity powering their servers to perform computations. The rest was essentially used to keep servers idling and ready in case of a surge in activity that could slow or crash their operations" [17]. The average utilization for hyperscale operators, such as

Google, is higher but has still historically been less than 50% [7, 33].

One way to reduce server idling is to put servers to sleep when they are not being used. The central concern with such a scheme is that latency involved with waking servers up to meet spiking loads will decrease quality-of-service. However, drawing an interesting parallel between computing loads in data centers and power loads on the electricity grid, Katz *et al.* [33] believe that a program of waking and sleeping servers based on a predictable "base load" while maintaining a "spinning reserve" to provide headroom for stochastic, bursty components of the load will work for many types of services. Other researchers [e.g., 68], are working on formal methods to optimize server provisioning for different load parameters. These sorts of methods could provide data center operators more confidence to sleep or shut down idling servers without risking service quality.

In addition to the designed overhead, there is also an issue of "comatose" servers—those that are no longer needed but are still running because no one can positively determine that the data is old or wants to risk pulling the plug: "[a]necdotal evidence indicates that 10-30% of servers in many data centers are using electricity but no longer delivering computing services. These servers have not yet been decommissioned and are probably not counted in installed base statistics. In many facilities nobody even knows these servers exist..." [23, p. 7]. Figures for comatose servers may be even higher. A sample at a LexisNexis data center revealed that three-fourths of installed servers used, on average, 10% of their capacity [17]. Both intentional overcapacity and failure to consolidate and decommission old equipment result in very low utilization rates.

In addition to retiring comatose servers, virtualization, colocation, and moving to the cloud have been shown to increase utilization rates. A recent industry survey revealed the following findings regarding migration to the cloud [69]:

- Large companies are more likely to pursue cloud computing than small companies, which is interesting because the cloud should be appealing to smaller entities that don't want the overhead associated with running a data center.
- Adoption of cloud computing is heavily skewed towards technology service providers. Traditional large vertical enterprises are much less likely to use the cloud.
- Companies reluctant to move to the cloud cite security, compliance, and reliability concerns.

Cloud migration is increasingly recognized as a major factor in reducing both energy use and emissions [8, 70, 71], and much of this benefit derives from the higher utilization delivered by virtualization cloud environments.

2.4.4 Power sourcing

The source of a data center's electricity has a large effect on its GHG and criteria air pollutant emissions, much like other industrial users. While Apple may not be focusing on driving down PUE to the extent that Google and Facebook are, the company emphasizes its clean energy sources for its data center electricity. If data center operators value a low carbon footprint, they must address the emissions of the power supply either by siting in areas where the grid has a low emissions factor, or by obtaining a separate source for cleaner power.

2.4.5 Prioritizing energy-efficiency improvements

Which of these four areas provide the most benefit? Masanet *et al.* [28] suggest that efficiency potential on the IT side (measured by ITUE and utilization) is larger than that on the infrastructure side (PUE). They also argue that, in an environment where renewable energy is limited from the grid, data centers may just be displacing other consumers of clean energy and should therefore focus on improving efficiency rather than on power sourcing.

Several reports have mentioned that utilization rate is more important than the efficiency of the equipment itself and is where the biggest immediate gains can be made [69, 71]. The NRDC report [71] compared on-premise data centers to the cloud environment with respect to carbon emissions. The model uses PUE, server utilization rate, server refresh period, virtualization ratio, and grid emissions factor estimated for several typical data center deployment scenarios. The study found that the most impact could be gained by targeting server utilization and electricity source emissions factor, and only then by improving infrastructure efficiency. Neither of these proposed focus areas are measured by PUE. Perversely, a decision to reduce energy consumption by shutting down idle servers in an existing data center would likely increase (worsen) the facility's PUE [59]. Virtualization and moving to the cloud significantly increase efficiency and reduce GHG emissions.

2.5 Conclusions and discussion

2.5.1 Toward a more holistic view of data center energy efficiency

Power Usage Effectiveness (PUE) has become an industry standard for reporting data center energy performance. While it is a useful measure of facility overhead, it is an incomplete metric. From an energy efficiency standpoint, it does not include the efficiency of the computing hardware. From a business standpoint, PUE does not measure energy productivity. Finally, from an environmental standpoint, it does not account for the carbon emissions associated with data center electricity use.

Based on these drawbacks, the industry's past focus on PUE is misplaced. While data center operators should of course continue to adopt best practices related to facility power and cooling, they would be better served by pursuing measures to increase utilization rates, reduce redundancy, and source clean power rather than continuing to push for marginally lower PUE numbers once the facility infrastructure is reasonably efficient. Indeed, the technology leaders in this industry are doing so.

Achieving data center efficiency requires revising the way these facilities are typically handled at the strategic level. Koomey and Flynn [72] advocate treating data centers as an "engine for cost reduction and competitiveness" instead of as a cost center, which requires a close link between IT metrics and business processes. A company that has integrated data center metrics with business operations—through metrics such as *energy per transaction*—will be more likely to identify and correct issues like low utilization, comatose servers, and inefficient facilities and hardware. Similarly, companies for whom a low carbon footprint is a strategic goal will naturally meet greater success by tracking data center carbon metrics. To be effective, the link between metrics and corporate goals should be supported by three "pillars" [73]: tracking the metrics in real time, establishing management and operational procedures, and using data center modeling and simulation software to continually test and optimize these procedures against the metrics to better meet the strategic goals.

2.5.2 Research issue: data availability

The lack of specific data about how servers are being configured and used as well as specific, benchmarked energy usage for certain types of equipment such as storage and network devices is a significant research issue. This lack of data is a recurring theme in the "future work" sections of literature and was confirmed in a conversation with EPA's products manager for the ENERGY STAR labeling for data center equipment. To deal with this want of data, two strategies are observed in the literature. Most energy use assessments with the goal of estimating total energy consumption use a bottom-up approach where estimates of server stocks, together with parameters like server utilization and PUE, and are fed into relatively simple calculations to render an estimate for overall sector energy consumption. Few studies, however, report uncertainty ranges on their results, though several parameterize different scenarios in their models.

Alternatively, some researchers [e.g., 71] look at the relative performance of different data center configurations. This approach eliminates the reliance on estimates for server stocks, though estimates for equipment performance, utilization rates, and other parameters are still necessary. These studies do not attempt to calculate absolute energy consumption, but rather show the magnitude of relative gains that can be made.

The first approach estimates energy use for the aggregate fleet but does not target energy efficiency, while the second evaluates potential efficiency gains for generic systems but does not provide insight on what is deployed and how it is used sector-wide. Aslan *et al.* [21] provides a summary of the different methods used to estimate the energy intensity of ICT infrastructure.

We believe the data gap could be addressed by establishing a framework under which data center operators report general data center characteristics and performance metrics. Though results from voluntary reporting would likely be skewed towards high-efficiency data centers, such results would still provide a broader window into the data center fleet than what is currently available publicly. Reported data could be sufficiently anonymized to alleviate security concerns while providing the industry and researchers alike a more complete picture of current performance. Indeed, several governments have programs in place for data center reporting, either under voluntary certification programs or carbon reduction mandates.

The U.S. EPA's ENERGY STAR voluntary data center certification program lists 73 certified facilities, though minimal information about each facility is reported [74]. The U.S. also has several programs to collect data on federally-owned data centers through the DOE Sustainability Performance Office and the Federal Data Center Consolidation Initiative, though these data are not publicly available. The National Australian Built Environment Rating System, which allows building operators to rate and certify their facilities, has a data center category, but only nine data centers are currently reported in the database

[75].

Under the UK's Carbon Reduction Commitment (CRC), large consumers of electric power, including data centers, are required to baseline and report electricity usage annually. The Environment Agency converts electricity usage to CO_2 emissions, which are reported for each firm [76]. Data centers are not easily distinguished from other types of facilities, nor are any other performance metrics or characteristics published. The European Commission's Joint Research Centre Institute for Energy and Transport has developed the European Code of Conduct for Energy Efficiency in Data Centres [77], a voluntary program. Participation in the code of conduct includes a reporting form with useful characteristic and performance data; however, the database of participants is not publicly available.

Given that lack of data seems to be a recurring complaint in the policy and research arenas, it would behoove these communities to come together with industry to establish a reporting framework. While measurement of data center performance has important nuances, and legitimate security and competitiveness concerns exist, the technical and institutional barriers to creating such a data set do not seem insurmountable.

Chapter 3

Known Unknowns: Indirect Energy Effects of Information and Communication Technology

A summary of this chapter has been published as Chapter 5 of A. Shehabi et al.*, "United States Data Center Energy Usage Report," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-1005775, Jun. 2016 [8]. A fuller version of this chapter is in review at* Environmental Research Letters.

> **Motivating questions: is there a consensus in the literature on the sign and magnitude of the net energy effect of ICT deployment? What are the mechanisms and dynamics through which ICT can either increase or decrease net energy use? What research priorities would help further the state of knowledge on indirect ICT energy effects?**

The rapid growth and adoption of information and communication technology (ICT) such as computers, mobile devices, sensors, and networks can profoundly affect how—and how much—energy is used by society. On the one hand, ICT itself consumes energy, which is a *direct energy effect.* On the other hand, ICT enables us to make existing products and services more efficient as well as create "e-substitutes" for physical products, which are indirect energy effects. Other, higher-order indirect energy effects occur when the introduction of ICT causes a change in consumption or production elsewhere in the economy.

The "digital revolution" has coincided with an increasing focus on environmental sustainability, and potential synergies between ICT deployment and efforts to mitigate environmental and climate impacts are popular topics of discussion among researchers, policymakers, and industry stakeholders. These synergies might be either positive or negative—ICT indirect effects could either offset or amplify direct energy effects—and characterizing this overall balance has been of great interest, as evidenced by the emergence of an *ICT for Sustainability* research community [78], two special issues in the *Journal of*

Industrial Ecology [79, 80] and one in *Resources, Conservation, and Recycling* [19], an OECD effort to link statistical indicators between the ICT and environment research fields [81], increasing work in "green computing" from the computer science literature [e.g., 82], and a variety of other reports [e.g., 83, 84]. Motivations behind ICT energy research are diverse: some seek a new carbon-abatement lever in the face of dire climate projections, while others hope to highlight the benefits of an industry often in the spotlight for its energy consumption.

There is, in fact, abundant cheerleading for ICT's ability to aid the cause of energy efficiency [e.g. 85, 86], even as there are rumblings about potential false promise [17]. Indeed, uncertainty regarding the magnitude and even the sign of ICT energy effects persists. Generally in the positive synergy camp are Romm *et al.* [87], Laitner and Ehrhardt-Martinez [88], the American Council for an Energy-Efficient Economy [e.g., 89] and the Center for Climate and Energy Solutions [e.g., 90], who anticipate ICT-enabled energy efficiency gains across broad sectors of the economy. The "SMARTer 2030" report [91], estimates a greenhouse gas (GHG) abatement potential of 20% by 2030 due to ICT deployment.[1] More cautionary is Rattle [92], who, in Chapters 5 & 6 of *Computing Our Way to Paradise?*, argues that higher-order effects are likely to swamp these sorts of energy savings projections.

In contrast, Berkhout and Hertin [93] argue for moving "beyond the dichotomy between pessimism and optimism" to recognize that the relationship between ICT and energy impacts is "complex, interdependent, deeply uncertain and scale-dependent." Other literature reviews point to an ambiguous net impact or acknowledge that this complexity and uncertainty confound attempts to verify a general belief that the net energy savings of ICT will be positive [94–97].

Our paper builds on this previous work. First, we answer the call made by Börjesson Rivera *et al.* [94] for standardization in the terms used across the literature by synthesizing the various published categorizations of ICT impacts into a common taxonomy (Section 3.1). Second, we review studies of individual ICT services—which constitute the bulk of the literature—providing a quantitative snapshot of the range of anticipated energy effects (Section 3.2). Next, we discuss higher-order energy impacts of ICT deployment, an area with much less solid quantitative treatment in the literature (Section 3.3). We conclude by summarizing the literature and highlighting directions for further research.

[1] Many, but not all, of the GHG emissions abatements result from decreased energy use.

3.1 Taxonomy of ICT energy effects

Direct energy consumption refers to energy used during the operation, manufacture, and disposal of ICT equipment. While this definition reflects common usage in the ICT energy literature, we note that it may contrast usage elsewhere—for instance, in economic input-output analysis—where direct energy use may be synonymous with operational energy consumption, and manufacturing and disposal energy are sometimes described as indirect effects [98]. Figure 3.1 shows past estimates and forecasts of ICT operational[2] energy consumption. The variation results from differing scopes (i.e., the equipment types included) and assumptions about equipment penetration, usage, and growth. For context, worldwide ICT direct operational electricity consumption has been estimated to be 655-710 TWh for 2007 and 905 TWh for 2012 [99, 100]. These site electricity estimates are generally on the order of 3-5% of total electricity consumption for their respective scopes.

In addition to the broad sector estimates in Figure 1, one subset of ICT operational energy use that has received careful study is energy consumption in data centers. After nearly doubling between 2000 and 2005, consumption growth is now nearly flat, having grown only 4% from 2010-2014 [8, 22, 108]. This reduction is driven by virtualization and consolidation of data center processing in “cloud” facilities and by increasing focus on energy-efficient data center IT infrastructure.

Energy consumption during other parts of the ICT equipment lifecycle—i.e., manufacture and disposal—is often called *embodied energy* and can be a nontrivial component of ICT equipment’s direct energy use. The relative significance of embodied energy to operational energy varies by component and by scope of analysis. Williams [109] observes that manufacturing energy accounts for well over half of the lifetime energy consumption for laptop computers and memory chips but less than 20% for logic chips.[3] At the data center level, Masanet *et al.* [28] estimate that operational energy dwarfs embodied energy.[4] The difference between laptop computers and data centers stems from the higher utilization rate of servers; to a lesser extent, the additional energy consumption of cooling needed in data center facilities also has an impact. At an even broader level, Raghavan and Ma [37] estimate that the embodied energy of the entire Internet infrastructure is roughly equivalent to its operational energy consumption

[2] I.e., not including manufacturing or disposal energy consumption.

[3] See Figure 2 in referenced paper and Figure 9 in Koomey *et al.* [96] for a comparison of embodied vs. operational energy for different ICT components and devices.

[4] Masanet *et al.* [28] report emissions, rather than energy, but their emissions estimates are derived from an energy model and the U.S. average fuel mix and are thus proportional to energy consumption.

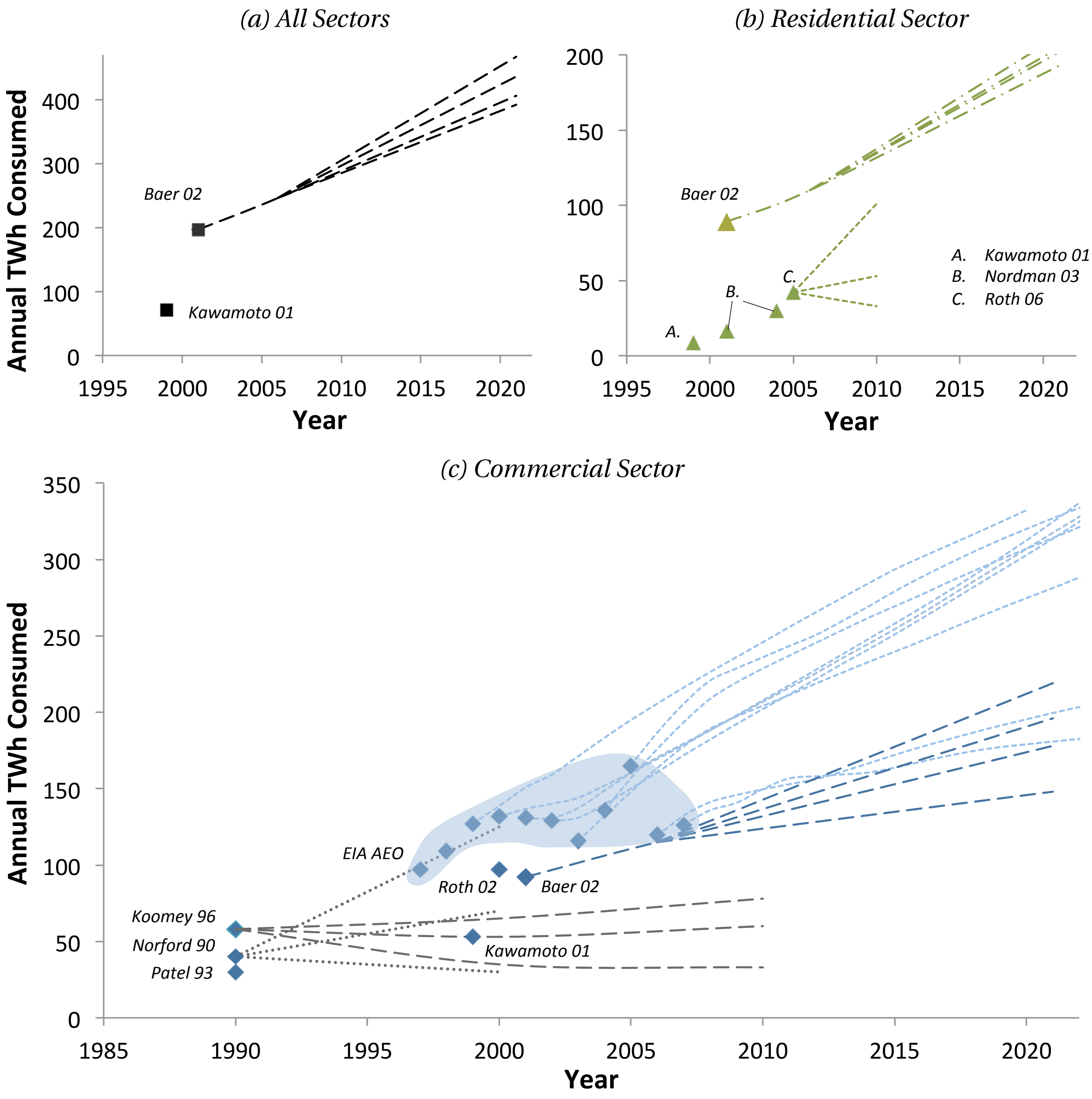


Figure 3.1: Estimates of use-phase ICT electricity consumption in the United States. Markers and solid lines represent historical estimates; dashed lines represent projections. Note different axis scales. The type of included ICT equipment varies significantly among the different studies.[a]

[a] AEO includes the EIA *PC office equipment* and *non-PC office equipment* categories, the latter including servers, copiers, fax machines, typewriters, cash registers, and other miscellaneous office equipment. Norford *et al.* [101] include PCs and their associated peripherals, including printers. Koomey *et al.* [102] include minicomputers, mainframes, point-of-sale terminals, fax machines, copiers, printers, monitors, and PCs. Kawamoto *et al.* [103] include portable computers, desktops, servers, displays, minicomputers, mainframes, terminals, laser and inkjet printers, copiers, and faxes. Roth *et al.* [104] includes PCs, servers, displays, copiers and printers, power supplies, and some computer and telephone networking equipment. Nordman and Meier [105] include desktop and laptop PCs, printers, copiers, and fax machines. Roth *et al.* [106] include PCs and peripherals (monitors, printers, and power supplies), multi-function devices, home networking equipment, set-top boxes, and broadband access devices. Baer *et al.* [107] takes the broadest view of ICT equipment, including TV and audio equipment in addition to office, networking, and communications equipment in the residential sector, and data centers in addition to office and networking equipment in the commercial sector.

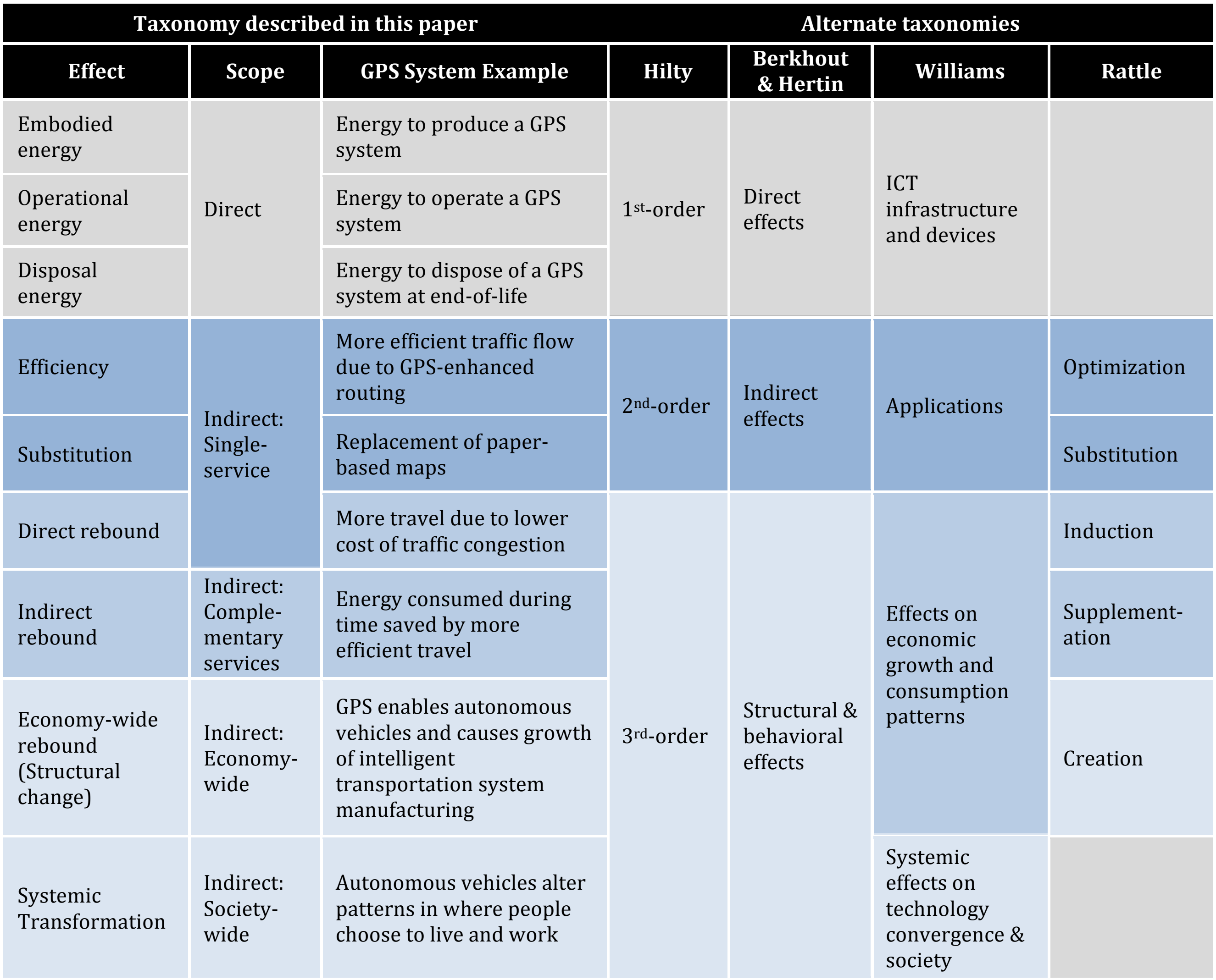
over its lifetime, which is partly due to the fact that network cabling has high embodied energy but no operational energy use.

However, direct energy use is likely the simplest and least important ICT energy effect [110]. The indirect energy effects are likely to be of much greater magnitude [96], owing to the breadth of the various mechanisms by which ICT services alter energy use. Furthermore, the electrical efficiency of computing has consistently doubled every 1.5 years [23], meaning that each kWh of direct energy use has the potential for ever-larger associated indirect effects. Table 3.1 breaks out individual effects, organizes them into a taxonomy of increasing scope (see also [111]), and maps them to other terms used in the literature, while Figure 3.2 shows this taxonomy graphically.

First, ICT adoption leads to *efficiency* in and *substitution* for conventional products and services. Efficiency improvement occurs when, for example, smart building technology reduces air conditioning energy consumption by tailoring climate-control to the real-time needs of building occupants. An example of substitution is the replacement of air travel with teleconferencing. There is no guarantee, however, that the substituted ICT service will be less energy intensive than the conventional service it replaces, and even evaluation of simple cases is not always straightforward. An electronic billboard, for instance, may use more energy than a static, printed billboard, since it uses electricity to display the image [112]. This energy consumption can be compared to the energy required to print the same image. However, the electronic version also avoids energy associated with changing the billboard—i.e., sending a worker out to make the switch. An additional complication is that the services are not strict functional equivalents: the electronic version allows animated displays, which may lead to higher success rates and profits—perhaps making energy consumption per successful "target" *less* even as per-billboard consumption is higher.

Any energy reduction achieved through efficiency or substitution can be plagued by *rebound effects*, in which expected gains are offset by induced additional consumption. Azevedo [111], Gillingham *et al.* [114], and Borenstein [115] provide comprehensive introductions to rebound effect types. Rebound is typically broken into direct rebound, indirect rebound, and economy-wide effects. *Direct rebound* effects are energy service own-price-elasticity effects: as prices fall (due to improvements in efficiency or productivity), substitution and income effects increase consumption. For an ICT example, if an e-book is less costly than a conventional book, then consumers might purchase more books. Direct rebound is constrained by saturation: there is a limit to the number of books people will buy, no matter how cheap

Table 3.1: Taxonomy of ICT energy effects. Scope of effect increases from top to bottom. The third column provides an example of each effect type related to the deployment of Global Positioning System (GPS) technology.[a]



Taxonomy described in this paper			Alternate taxonomies			
Effect	**Scope**	**GPS System Example**	**Hilty**	**Berkhout & Hertin**	**Williams**	**Rattle**
Embodied energy	Direct	Energy to produce a GPS system	1st-order	Direct effects	ICT infrastructure and devices	
Operational energy		Energy to operate a GPS system				
Disposal energy		Energy to dispose of a GPS system at end-of-life				
Efficiency	Indirect: Single-service	More efficient traffic flow due to GPS-enhanced routing	2nd-order	Indirect effects	Applications	Optimization
Substitution		Replacement of paper-based maps				Substitution
Direct rebound		More travel due to lower cost of traffic congestion	3rd-order	Structural & behavioral effects	Effects on economic growth and consumption patterns	Induction
Indirect rebound	Indirect: Complementary services	Energy consumed during time saved by more efficient travel				Supplementation
Economy-wide rebound (Structural change)	Indirect: Economy-wide	GPS enables autonomous vehicles and causes growth of intelligent transportation system manufacturing				Creation
Systemic Transformation	Indirect: Society-wide	Autonomous vehicles alter patterns in where people choose to live and work			Systemic effects on technology convergence & society	

[a] Alternate taxonomies are from Rattle [92], Berkhout and Hertin [93], Williams [109], and Hilty *et al.* [113].

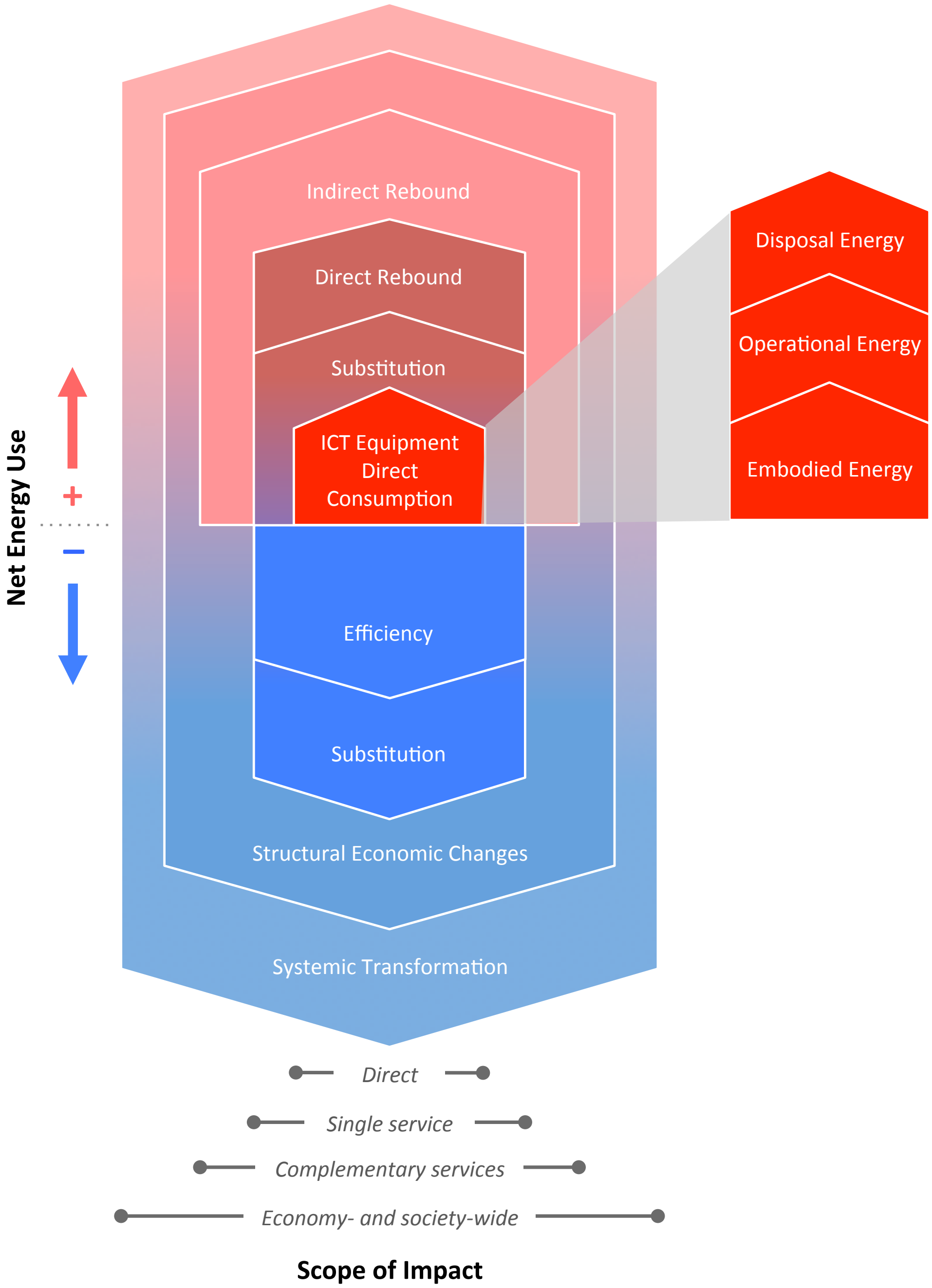


Figure 3.2: Taxonomy of ICT energy effects. Red effects increase energy use, blue effects decrease energy use, and shading intensity decreases as effect scope increases. (Effect magnitudes are only illustrative and not to scale.)

they become. Alternatively, these savings could be spent on other goods and services, which are *indirect rebound* effects. Indirect rebound effects result from cross-price elasticity of demand for other products and services due to increased real consumer income.[5]

Economy-wide effects occur when the ICT introduction causes macroeconomic adjustments across economic sectors. That is, the ICT industry can promote or inhibit growth in other sectors of the economy, inducing structural changes that have energy use implications of their own. For example, e-commerce is having broad effects on the logistics industry [116], including growth in urban freight vehicle sales and changing patterns in distribution center floor space [117], increased trucking and adoption of new pricing strategies by freight carriers [118], and use of more specialized packaging and a broader range of box sizes [119].

Finally, *transformational effects* refer to the altering of human preferences and economic and social institutions caused in part by the development of ICT [120, 121]. Historical examples include the advent of the telephone and automobile, which heavily altered where and how people lived and worked. We might conceive of a similar transformation (one of many possible ICT-enhanced futures) in which the fundamental constraints on where people live and work continue to loosen: e-commerce and home delivery make proximity to traditional retail outlets less important, seamless telework results in less commuting, and driverless vehicles allow for more productive use of the commuting time.

As noted by Börjesson Rivera *et al.* [94], the existing literature uses several different sets of terms for this hierarchy of effects. The right half of Table 3.1 maps the most commonly used categorizations to the taxonomy used in this chapter.

ICT energy effects can be broadly grouped into first-order impacts due to direct consumption, second-order effects resulting from process changes, such as efficiency, and third-order effects due to behavioral and economic changes [93, 113]. Williams [109] adds a fourth level, essentially breaking third-order effects into rebound effects and broader systemic change.

Rattle [92] categorizes indirect effects into five categories: optimization, substitution, induction, supplementation, and creation. The first two map directly to efficiency and substitution, while induction, supplementation, and creation align loosely with (or, perhaps more strictly, are special cases of) direct, indirect, and economy-wide rebound effects, respectively.

[5] Note that direct and indirect *rebound* do not correspond to the distinction between direct and indirect *energy effects* used in this and other papers; all rebound effects are indirect energy effects. See first two columns of Table 3.1.

3.2 Indirect single-service effects

Though it is important to take a systemic, holistic view of ICT energy consumption [110], tractability concerns dictate that researchers attempting detailed quantitative estimates of energy impacts look at specific applications separately. These granular studies, which often use a life-cycle assessment (LCA) approach, can identify the key factors driving energy use and highlight opportunities for reduction in individual processes. However, they do so at the expense of scope, typically addressing only substitution and efficiency effects. Bull and Kozak [122] discuss the challenges of LCA specific to the ICT domain. For a sample evaluation of ICT-related LCA studies, see Schmidt and Pizzol [123].

In this section, we review literature estimating energy consumption impacts attributable to the introduction of four ICT services—e-commerce, e-materialization, telework, and monitoring and controls—across the building, transport, manufacturing, packaging, and waste sectors. (See Figure 3.3.) These four services were selected due to their broad impacts and coverage in the literature, but there are other energy-relevant ICT services, such as computer-aided design (CAD), which has expanded beyond drafting software to cover process planning, engineering, and quality control [124]. Furthermore, increased computing power has enabled system designers to solve more complex problems using optimization, modeling, and simulation techniques and thus more comprehensively cover the "solution space," yielding products with greater function, lower cost, less embodied energy, and increased use-phase efficiency [124].

The aerospace industry provides a particularly clear example of the evolution of engineering design from manual methods to reliance on computational modeling and simulation. ICT has transformed all levels of aircraft design. First, the ability to solve complex design optimization problems supports development and use of new materials as well as enhanced design of aircraft components, such as airplane wings [125, 126]. In particular, multidisciplinary design optimization allows joint consideration of structures and aerodynamics in the design process [127, 128]. These efforts lead to both reduced material use in production as well as increased efficiency in flight. Second, ICT has made systems integration throughout the engineering and production process more efficient [129]. Commoditization and outsourcing of components can increase the energy efficiency of production, although potential increases in the transport involved in a global supply chain may increase energy use [130]. Finally, ICT has replaced wind-tunnel testing and even some flight testing, decreasing the manufacturing and embodied energy

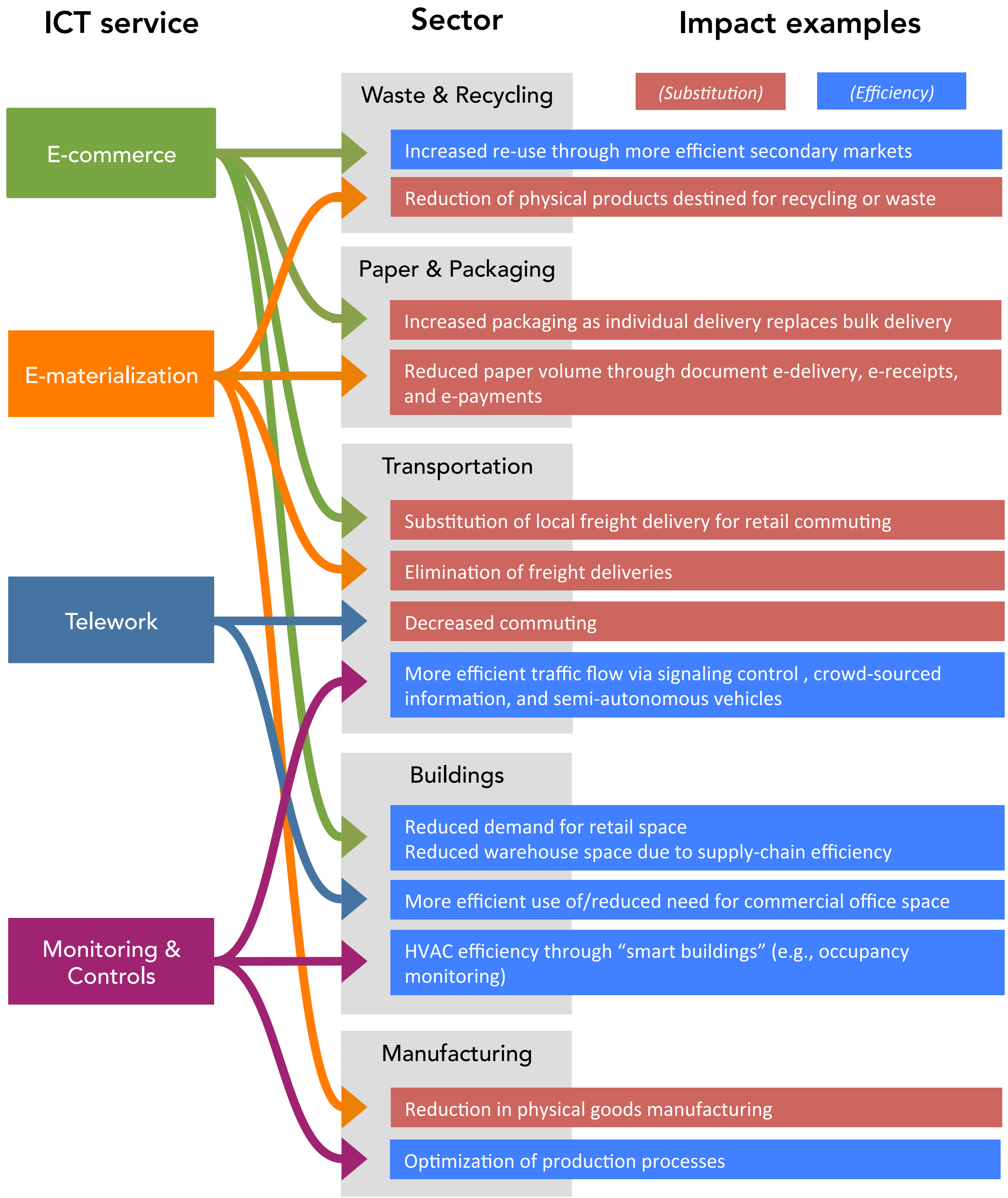

ICT service
Sector
Impact examples
(Substitution)
(Efficiency)
E-commerce
E-materialization
Telework
Monitoring & Controls
Waste & Recycling
Increased re-use through more efficient secondary markets
Reduction of physical products destined for recycling or waste
Paper & Packaging
Increased packaging as individual delivery replaces bulk delivery
Reduced paper volume through document e-delivery, e-receipts, and e-payments
Transportation
Substitution of local freight delivery for retail commuting
Elimination of freight deliveries
Decreased commuting
More efficient traffic flow via signaling control , crowd-sourced information, and semi-autonomous vehicles
Buildings
Reduced demand for retail space
Reduced warehouse space due to supply-chain efficiency
More efficient use of/reduced need for commercial office space
HVAC efficiency through "smart buildings" (e.g., occupancy monitoring)
Manufacturing
Reduction in physical goods manufacturing
Optimization of production processes

Figure 3.3: Relationships among ICT service types, economic sectors, and impacts.

of physical prototypes and reducing fuel use [131].

Similar effects could doubtless be found in other manufacturing or material-intensive industries, including consumer goods, automobiles, and construction [124, 132].

3.2.1 E-commerce

E-commerce, the buying and selling of goods and services using electronic networks, includes familiar business-to-customer (B2C) Internet outlets like eBay and Amazon, but it also includes back-end business-to-business (B2B) functions such as services that enable just-in-time inventory management. Though focused on GHG emissions, Table 1 in Siikavirta *et al.* [130] outlines different means by which e-commerce affects energy consumption throughout the supply chain.

A review of e-commerce studies, summarized in Table 3.2, shows mixed results. On balance, most studies find a positive potential energy savings, though this conclusion is not universal, and results are highly sensitive to assumptions [116]. The series of studies examining book retail is instructive on this point, since these analyses were completed by the same research community[6] using similar methods with similar (though not identical) system boundaries.

In the transport sector, a switch from brick-and-mortar retail to electronic retail changes how products are delivered to the consumer, with personal travel and bulk freight delivery to stores replaced by home delivery. E-commerce may make "last mile" transport more efficient due to optimization of shipping routes by delivery companies, but it can increase energy use by substituting air for ground freight. It also lowers package density, since traditional stores receive multiple items in each box, while home delivery entails fewer items per box, leading to higher embodied packaging energy [134]. Additionally, the long reach of e-commerce gives retailers the capacity to serve geographically larger markets, which could increase cost-efficiency at the expense of energy-efficiency. Most e-commerce studies focus on these transport and packaging effects. Among those in Table 3.2, key sensitivities driving results are population density (related to last-mile delivery), freight mode, product return rate, trip allocation (proportion of multipurpose trips), and packaging type.

As an example of how system assumptions affect results, we highlight the *negative* 500% net savings (that is, a 5x increase in energy consumption) from Matthews *et al.* [135]. The primary driver of results

[6] With the exception of Romm *et al.* [87] and Kim *et al.* [133].

in this study is transport distance, which is a function of population density. This particular estimate reflects the high-density Tokyo scenario, in which customers live within half a kilometre of a bookstore and are thus likely to walk or ride a bicycle when shopping. In the e-commerce case, courier trucks are used for delivery. As a result, e-commerce requires ten times as much total transport energy compared to traditional retail. The Tokyo result is, of course, an outlier when compared to the U.S. scenarios and the other Japan scenarios in the same paper. However, it represents a valid model of the system and is thus a particularly clear—if extreme—example of how results are driven by the assumed characteristics of the system.

In the buildings sector, Romm *et al.* [87] estimate a potential for 53 billion kWh in operational and construction energy reductions in retail, warehouse, and office space due to B2C e-commerce from 1997 to 2007. Mechanisms for achieving this reduction primarily include shrinkage, consolidation, or replacement of brick-and-mortar retail outlets but also, e.g., more efficient use of hotel rooms through Internet bookings and auctions. In the B2B segment, they estimate supply-chain efficiency will reduce inventories by 25% to 35%, leading to elimination of 1 billion square feet of warehouse space from 1995 levels. Matthews and Hendrickson [143] find a net reduction in logistics energy use through the centralization of inventory, much of which is likely enabled by ICT.

Through greater coupling between consumers and producers, e-commerce may reduce overproduction. E-commerce also leads to more efficient secondary markets. Through sites like eBay, Craigslist, and Freecycle, goods that were either destined for the landfill or sitting unused in storage are put to use, eliminating waste, avoiding some manufacturing, and reducing storage requirements. At the same time, these secondary markets can increase energy consumption, specifically in transport [92, p. 71].

3.2.2 E-materialization

In addition to altering delivery channels for physical products, Internet-based retail allows for the substitution of some products with electronic equivalents, i.e., *e-materialization*, *virtualization*, or *digitization*. Consumer examples include electronic vs. print newspapers, e-books vs. bound books, and streaming vs. physical media such as CDs and DVDs. In business operations, e-materialization can lead to reduction in paper communications and records. The theoretical energy impacts of e-materialization across the transport, manufacturing, packaging, and waste sectors are straightforward: elimination of physical products eliminates the need to manufacture, package, transport, and dispose of those products. Offsetting these gains is the direct energy consumption of the ICT used to deliver the virtual substitutes.

Results from e-materialization studies are summarized in Table 3.3. Online media streaming (vs. shipping CDs/DVDs by mail) is a popular e-materialization use case, and comparison of results for this service highlights the variability common to LCA studies, even when the dynamics of the service are well known and fairly straightforward. Additionally, Bull and Kozak [122] argue that the inherent complexity and interconnectedness of ICT systems weaken LCA's ability to provide meaningful comparative results.

Key assumptions driving this variability include energy consumption by the network and end-user devices, media file size, and media re-use; the electronic delivery option becomes less competitive as network energy, file size, and frequency of re-use increase.

As we did above, we highlight an example study with wide-ranging results. Gard and Keoleian [144] investigate six different scenarios comparing electronic and paper library journals, finding savings ranging from a 643% increase to a 69% savings in energy use. The large increase in energy use for digital journals occurs in a scenario in which each article is read a thousand times (spread across 100 different libraries). Multiple readings skew the results in favor of paper journals, since each read beyond the first is essentially free, whereas each read of an electronic copy incurs ICT energy consumption. However, subsequent scenarios added printing and copying of articles and personal transport to and from the library, which reduced the advantage of the traditional publication. The 69% savings occurred when readers drive to the library to read the paper copy but can access the digital copy from home. Clearly, some of the scenarios are less reflective than others of how the journal publication system exists today; yet, fifteen years ago, these conclusions identified factors which could inform the evolution of this system. For instance, providing library patron access to journal articles from home not only increases convenience, but can flip the net savings effect for this service from negative to positive—in some sense, rendering concerns about direct ICT consumption of this service moot.

3.2.3 Telework

Telework refers to the use of virtual collaboration and teleconferencing software, networks, and electronic file systems to enable employees to work remotely from an alternate location. Telework can potentially reduce energy used in personal transport as employees avoid commuting by working from home and as face-to-face meetings are replaced by teleconferencing. In the buildings sector, home offices might increase residential energy consumption while decreasing commercial consumption through higher utilization of existing offices (through space-sharing) and avoided new construction.

Table 3.4 summarizes estimates of these energy effects. Varying greatly in method and scope, the telework studies do not lend themselves to comparing quantitative results, so we report findings specific to each study rather than savings percentages. Approximately half of the studies are optimistic about energy savings, while the other half are more guarded—either finding savings to be modest in the overall energy picture or finding that savings can be positive or negative depending on parameters. The most important driver of savings is frequency of teleworking; infrequent telecommuters may cause a net increase in energy use due to redundancy in home and central offices, whereas regular telecommuters allow for larger reductions in commercial consumption.

Importantly, while a few of these studies do incorporate some aspects of direct rebound—usually by acknowledging that personal errands usually combined with the work commute must be undertaken separately—broader rebound considerations are not included, and thus these results may be optimistic. Conversely, Aebischer and Huser [149] note a reason for net benefits being *underestimated*: the definition of teleworking in most studies excludes those workers for whom ICT enables self-employment.

3.2.4 Monitoring and controls

ICT has increased the frequency and precision with which we are able to monitor and control energy-consuming processes, enabling a higher degree of process optimization. Table 3.5 summarizes a wide range of studies across the transport, buildings, and manufacturing sectors. While the energy savings are positive in most of these studies—being, as they are, focused on efficiency—most do not account for the direct ICT energy use, and so the net savings will be less than reported.

ICT deployment in the transport sector is broad and multiscale. Focusing our discussion on road vehicles, components like the fuel injectors and throttle are monitored and controlled in real time to optimize fuel economy and provide fault-detection alerts; at the system level, networked vehicles and road infrastructure sensors monitor traffic, enable rerouting, and inform variable speed limits. Route optimization studies find fuel savings on the order of 10%, with additional savings of 1-4% achievable through utilization of route information in adaptive drivetrain control. Other ICT-enablers include weigh-in-motion sensors (reducing truck diesel consumption), car- and ride-sharing (reducing urban car ownership), and real-time bus tracking (increasing appeal or convenience of public transport). Brown *et al.* [155] provide a comprehensive review of various vehicle automation technologies and summarize literature estimates of eleven energy effects each might yield. Langer and Vaidyanathan [156] describe the ways in which ICT-enabled “smart-freight” can reduce energy use in cargo transport.

Turning to buildings, Meyers *et al.* [157] estimate that average U.S. residences waste around 40% of their primary[7] energy consumption due to inefficiency. Much of this waste is addressable by ICT interventions. Smart meter technology coupled with displays can provide real-time load information, which should cause a rational (in the classical economic sense) customer to reduce consumption. However, many studies find underwhelming savings from smart metering [158], and such studies may be biased [159] or confounded by the Hawthorne effect, in which participants alter their behavior simply because they are aware that the study is taking place [160].

Building energy management systems (BEMS), including technology like programmable thermostats and occupancy sensors, can reduce the need for human hands (and minds) to make routine energy-saving interventions. BEMS match heating, ventilating, and air conditioning (HVAC) operation to required load and analyze consumption patterns to detect faults. Empirical studies of BEMS have found energy savings of 7-23%. Rogers *et al.* [161] estimate reductions of between $37 and $85 billion in annual energy costs by “intelligent efficiency” technologies in the commercial and manufacturing sectors by the year 2035.[8] Aebischer and Huser [149] express some concern over both rebound and direct consumption, positing that those installing advanced lighting control technology may be more likely to also wire more lights and noting that standby consumption for such systems is also higher. The advent of low-power sensors and controllers [96] may mitigate this last concern.

In manufacturing, industrial control systems increase efficiency, fault-detection, and productivity, reducing per-unit energy consumption and wastage [93, 107]. The vision for achieving this potential energy savings through *smart manufacturing* is laid out in the DOE’s 2015 Quadrennial Technology Review [162] and a European Commission report [124]. Actual savings results are hard to tease out, as modern manufacturing processes are already heavily integrated with ICT, with much of the publicly-available

[7] The authors conduct their analysis on primary energy—i.e., “inclusive of energy use upstream in the fuel cycle.”

[8] See Table 4 in the reference for a list of savings ranges for specific energy efficiency technologies in the commercial sector.

insight coming from DOE case studies [162, 163]. Nonetheless, ICT is a key enabler in energy-efficient manufacturing [164, 165], and industry stakeholders are emphasizing ICT-enabled efficiency over the next decade [166, 167].

ICT monitoring and control has also proved beneficial in the power sector, enabling more aggressive demand-side management (DSM); however, many DSM programs simply shift use to reduce peak loads rather than avoid the energy use overall [168]. If, however, such load shifting ultimately avoids the construction of power plants or deployment of diesel generators, then the embodied energy of that infrastructure is saved.

Table 3.2: Summary of e-commerce studies. Net savings is energy savings of ICT service vs. conventional baseline. Where point estimates rather than ranges are provided, the value is placed in the *High* column, though it may be an average. *Qualitative conclusion* is an assessment of where the bulk of the evidence in the study falls; ▲/▼ indicate positive and negative ICT energy savings, respectively; ◊ represents a balanced finding (i.e., savings offset or are balanced between positive and negative results depending on parameters). Assignment of these icons is based on the original authors' results discussion in each paper, augmented by our interpretation.

Study	Service	Region	Effects							Sectors					Metrics		Net Savings			Method
			Direct	Efficiency	Substitution	Direct Rebound	Indirect Rebound	Economy-wide rebound	Transformation	Transport	Buildings	Manufacturing	Packaging	Waste	Air Emissions	Energy	Low	High	Qualitative conclusion	*(LCA=Life Cycle Assessment, EIO=Economic Input-Output, MC=Monte Carlo, SD=System Dynamics)*
Siikavirta *et al.* [130]	Food retail	Finland		■						■					■		18%	87%	▲	Simulation
Romm *et al.* [87]	Book retail	US	■	■							■					■		93%	▲	Calculation
Matthews *et al.* [136]	Book retail	US		■						■		■	■		■	■		16%	▲	LCA (EIO)
Matthews *et al.* [137]	Book retail	US		■						■		■	■		■	■	-7%	9%	◊	LCA (EIO)
Matthews *et al.* [135]	Book retail	US		■						■		■	■			■	-32%	18%	◊	LCA (EIO)
Matthews *et al.* [135]	Book retail	Japan	■	■						■	■		■			■	-500%	28%	◊	LCA (Process)
Williams and Tagami [134]	Book retail	Japan	■	■						■	■		■			■	-51%	44%[a]	▼	LCA (Process)
Kim *et al.* [133]	Book retail	US		■						■					■	■		51%	▲	Simulation
Sivaraman *et al.* [138]	DVD rental	US	■	■						■	■	■	■		■	■	23%	50%	▲	LCA (Hybrid)
Shehabi *et al.* [139]	DVD rental	US	■	■						■	■	■	■		■	■		35%	▲	LCA (Hybrid)
Weber *et al.* [140]	Music retail	US	■	■						■	■	■	■		■	■	-97%[b]	71%	◊	LCA (Process) w/ MC
Erdmann *et al.* [83]	"Tele-shopping"	EU-15	■	■		■				■	■	■	■	■	■	■		1%	◊	Simulation w/ SD
Weber *et al.* [141]	Consumer retail	US	■	■						■	■		■		■	■	< 0[c]	36%	▲	LCA (Process) w/ MC
Edwards *et al.* [142]	Consumer retail	UK		■						■					■			n/a[d]	▲	Simulation

[a] Derived from seeking the minimum value for e-retail. Main scenario results were all negative.

[b] This study does not report the scenario *differences* from the Monte Carlo runs, so the bounds shown here are the maximum possible positive and negative savings based on the confidence intervals reported in the study. It is likely, however, that there is correlation among the scenarios and the range is not this large. Median savings estimates are reported as 20%-30%.

[c] The study does not report results in enough detail to determine the full range of values, but cites a 20% probability that the traditional channel has lower energy use than the e-commerce channel.

[d] Last mile transport only, so results not comparable to other studies in this table. Generally, e-commerce had much lower per-item energy use in this study.

Table 3.3: Summary of e-materialization studies. See Table 3.2 caption for information on symbols.

Study	Service	Region	Effects							Sectors					Metrics		Net Savings			Method
			Direct	Efficiency	Substitution	Direct Rebound	Indirect Rebound	Economy-wide rebound	Transformation	Transport	Buildings	Manufacturing	Packaging	Waste	Air Emissions	Energy	Low	High	Qualitative conclusion	*(LCA=Life Cycle Assessment, EIO=Economic Input-Output, MC=Monte Carlo, SD=System Dynamics)*
Seetharam *et al.* [145]	Video delivery	US	■	■	■					■	■	■	■	■	■	■	21%	70%	▲	LCA (Process)
Shehabi *et al.* [139]	Video delivery	US	■	■	■					■	■	■	■		■	■		-1%	◊	LCA (Process)
Weber *et al.* [140]	Music delivery	US	■	■	■					■	■	■	■		■	■	-30%	90%	▲	LCA (Process) w/ MC
Mayers *et al.* [146]	Game delivery	UK	■	■	■					■	■	■	■		■		-32%	-5%	▼	LCA (Process)
Moberg *et al.* [147]	News media	EU	■	■	■					■	■	■		■	■	■		60%	▲	LCA (Process)
Erdmann *et al.* [83]	Virtual goods	EU-15	■	■	■	■				■	■	■	■	■	■	■	0%	11%	▲	Scenarios w/ SD
Gard and Keoleian [144]	Library journals	US	■	■	■					■	■	■		■		■	-643%	69%	◊	LCA (Process)
Zurkirch and Reichart [148]	Mail delivery	Switzerland	■	■	■					■	■	■	■	■	*[a]	*	-80%	0%	▼	LCA (Process)
Zurkirch and Reichart [148]	Phone book	Switzerland	■	■	■					■		■	■	■	*	*	0%	93%	▲	LCA (Process)

[a] Zurkirch and Reichart [148] use Ecopoints, an LCA metric that is a weighted combination of a suite of environmental effects.

Table 3.4: Summary of telework studies. See Table 3.2 caption for information on symbols.

Study	Region	Effects							Sectors		Metrics		Net Savings	Method	
		Direct	Efficiency	Substitution	Direct Rebound	Indirect Rebound	Economy-wide rebound	Transformation	Transport	Buildings	Air Emissions	Energy		Qualitative conclusion	*(LCA=Life Cycle Assessment, EIO=Economic Input-Output, MC=Monte Carlo, SD=System Dynamics)*
Romm *et al.* [87]	US	■	■	■						■		■	Savings = 1.5% of residential & commercial electricity	▲	Calculation
Aebischer and Huser [149]	Germany			■	■				■			■	24% reduction in vehicle travel	▲	Empirical survey
Aebischer and Huser [149]	Switzerland	■	■	■						■		■	-115 to 282 kWh/y/telecommuter, saved depending on frequency	◊	Case study
Baer *et al.* [107]	US	■	■	■						■		■	32 TWh electricity saved in 2001; 48-216 TWh by 2021[a]	◊	Scenario analysis
Atkyns *et al.* [150]	US			■	■				■			■	5.1 million gals. gasoline saved over 68K employees for 1 year	▲	Empirical survey + calculation
Hopkinson and James [151]	UK	■	■	■	■				■	■		■	0-50% commercial space saved; Commute decrease; business travel inconclusive	▲	Case study
Erdmann *et al.* [83]	EU-15	■	■	■	■				■	■	■	■	Telework & virtual meetings energy savings 1%	▲	Scenarios w/ SD simulation
Matthews and Williams [152]	US, Japan	■	■	■					■	■		■	0.01-0.4% net national energy savings	◊	Calculation
Roth *et al.* [153]	US	■	■	■	■				■	■	■	■	7-80 MJ annual savings per telecommuter	▲	LCA (hybrid)
Kitou and Horvath [154]	US	■	■	■	■				■	■	■	■	Avg direct energy cost savings of 18%	▲	Simulation w/ MC

[a] The study does not break out the proportion of ICT direct energy use allocated to solely telework applications, so the net effect is ambiguous. These are efficiency and substitution savings due to telework, without deducting increases in direct energy use. Overall ICT impacts in this study (for teleworking and other services) vary based on scenario.

Table 3.5: Summary of monitoring and controls studies. See Table 3.2 caption for information on symbols.

Study	Service	Region	Effects						Sectors			Metrics		Net Savings			Method
			Direct Efficiency	Substitution	Direct Rebound	Indirect Rebound	Economy-wide rebound	Transformation	Transport	Buildings	Manufacturing	Air Emissions	Energy	Low	High	Qualitative conclusion	*(LCA=Life Cycle Assessment, EIO=Economic Input-Output, MC=Monte Carlo, SD=System Dynamics)*
Ganti *et al.* [169]	Vehicle routing	Urbana, IL	■						■				■	6%	13%	▲	Experiment + calculation
Ericsson *et al.* [170]	Vehicle routing	Lund, Sweden	■						■				■		8%	▲	Simulation
Gonder [171]	Vehicle drivetrain	n/a	■						■				■	2%	4%	▲	Simulation
Huang *et al.* [172]	Vehicle drivetrain	US	■						■				■	1%	3%	▲	Simulation
Brown *et al.* [155]	Vehicle automation	US	■	■	■			■	■				■	-173%	95%	◇	Kaya Identity estimation
Erdmann *et al.* [83]	Intelligent Transport	EU-15	■	■	■				■			■	■		-3%	▼	Scenarios w/ SD simulation
Erdmann *et al.* [83]	BEMS	EU-15	■	■	■					■		■	■	4%	6%	▲	Scenarios w/ SD simulation
Erdmann *et al.* [83]	Supply chain/ process mgmt	EU-15	■	■	■						■	■	■	0%	2%	▲	Scenarios w/ SD simulation
Meyers *et al.* [157]	Residential energy	US	■							■			■	33%	62%	▲	Calculation
Mattern *et al.* [158]	Smart meters	n/a	■							■			■	2%	4%	▲	Meta-analysis
Davis *et al.* [159]	Smart meters	n/a	■							■			■	1%	3%	▲	Meta-analysis
Doukas *et al.* [173]	BEMS	Greece	■							■			■		10%	▲	Case study
Agarwal *et al.* [174]	HVAC	San Diego, CA	■							■			■	10%	15%	▲	Pilot deployment
Agarwal *et al.* [175]	HVAC	San Diego, CA	■							■			■	8%	16%	▲	Pilot deployment
Henderson and Waitner [176]	BEMS	DC	■							■			■	7%	23%	▲	Case study
Rogers *et al.* [161]	BEMS & Ind. Process	US	■							■	■		■	5%	75%	▲	Literature review
Masanet [163]	Industrial controls	US	■								■	■	■	50–3000 MMBTU/yr[a]		▲	Empirical data analysis

[a] per small or medium sized manufacturer

3.3 Indirect complementary, economy-wide, and systemic effects

Service-specific studies like those discussed above can highlight individual pathways for ICT to alter energy consumption, but they rarely address the higher-order effects beyond efficiency and substitution. These rebound and systemic effects are crucial to an integrated picture of whether—or under which conditions—ICT services lead to a net increase or decrease in system-wide societal energy use. If obtaining conclusive results for a particular service can be complex and uncertain, the macro picture is even more so.

The inability to draw concrete conclusions reflects, in large part, uncertainty regarding the rebound effect for ICT and the inability to disentangle root causes of interrelated economic effects. The dynamics of these effects are hugely dependent upon human behavior, which is laden with uncertainty and confounds efforts to achieve the full technical potential of ICT interventions [177].

While rebound could apply to all the services discussed in Section 3.2, telework-related rebound seems to have garnered the most interest in the literature—though a similar discussion surrounding more recent innovations such as ride-sharing services and self-driving cars is also emerging. Matthews and Williams [152] note that indirect effects from telework are likely much larger than the energy savings from substitution, though the sign could be positive or negative depending on which type of effect dominates. Mokhtarian [178] notes that travel in the U.S. has increased over the decades during which ICT might have been expected to reduce it and explains a number of ways, ranging from direct rebound through systemic change, in which ICT could stimulate more travel. For instance, ICT might generally decrease the personal cost of travel by making transportation more efficient, cheaper, and productive, and people might respond by commuting farther or taking more vacations. The model in Erdmann *et al.* [83] presumes a priori that this phenomenon of actually increasing energy use via large rebound—known as backfire [179]—will exist for intelligent transportation systems.

However, it should be noted that these studies rely either on conjecture or speculative models rather than empirical results. Indeed, empirical data are hard to come by, with a recent review of ICT rebound studies highlighting a literature heavier on conjecture and discussion than on results [180]. A set of case studies conducted in the EU find ICT-related rebound effects from e-commerce and telework ranging from 14% to 73% [84, Table 3.10]. Yet, even the careful attempt to base model parameters on empirical findings found in Mokhtarian [181] demonstrates the uncertainty in assigning parameter values and the high sensitivity of the results to these assumptions.

Another way of estimating rebound effects is to analyze macroeconomic data. Choo *et al.* [182] attempt to assess the aggregate effect of telecommuting on vehicle miles travelled (VMT) empirically using an econometric time-series model on macroeconomic variables. They find that telecommuting has had a statistically significant reduction on VMT, but noted that the average magnitude of this reduction estimated by the model seems implausibly large. They rely on external evidence to argue that the actual reduction is at the low end of their estimated confidence interval.

This macroeconomic approach can be applied to the overall ICT net effect as well. A batch of studies conducted near the turn of the century note that several years of accelerated decrease in energy intensity of the U.S. economy might in part be attributed to both structural and efficiency changes caused by ICT adoption [87, 183, 184]. Several economic studies test this sort of hypothesis using regression—i.e., estimating coefficients representing the relationship between energy consumption and explanatory variables like ICT investment, gross domestic product, and population. Laitner and Ehrhardt-Martinez [88] estimate that each kWh of direct electricity consumption by ICT equipment is responsible for between 6 and 14 kWh in energy savings in the US through efficiency and substitution. Other econometric studies of the U.S. and Japan find that ICT investment has led to decreased energy intensity [185, 186], though the latter study suggests that, as developed economies complete their energy-reducing transition from “smokestack” industries to ICT industries, ICT will eventually lead to increasing energy use due to the economic income effect.

In a sector-by-sector analysis of energy trends, however, Murtishaw and Schipper [187] attribute decreasing energy intensity of the U.S. economy from 1988-1998 to structural economic changes rather than efficiency gains, and they are unable to conclude that these structural changes resulted from ICT. Cho *et al.* [188], looking at individual sectors of the South Korean economy, find mixed results. Furthermore, Koomey *et al.* [96] note that economic models are generally not good at assessing situations where the structure of the economy is undergoing rapid change and that disentangling and attributing broad effects are difficult.

Other researchers use scenario analysis to consider sets of plausible alternate pathways, rather than trying to model existing dynamics. The hope is that thinking about how possible futures could unfold will make us better prepared to monitor and direct progress. Possible energy benefits highlighted in ICT futures include information access promoting “environmental literacy” in consumers [93, 189], innovation and agility in businesses [190, 191], and the easier integration of distributed and renewable generators on the electric grid [107].

A 2002 RAND report uses scenarios to assess possible ICT impacts on electricity consumption through 2021 [107]. The report includes estimates of direct ICT electricity use, efficiency gains resulting from building energy management systems (BEMS) and industrial process controls, and indirect effects from telework and e-commerce. (Since the report focuses on electricity use, these effects generally do not include impacts in the transport sector.) The ICT-driven electricity effect in the year 2021 ranges from *negative* 203 TWh to *positive* 200 TWh across the four scenarios.

Hilty *et al.* [113] take a scenario-based approach for Europe, though the study looks at other environmental impacts beyond energy. Unlike the RAND study, although there is uncertainty in overall future energy consumption (with total energy consumption in Year 2020 increasing by 37% in the worst-case scenario but decreasing by 17% in the best case), the expansion of ICT universally decreases overall energy consumption vs. the counterfactual base case where the level of ICT deployment remains constant. This decrease was small in aggregate, which the authors explain is the result of ICT-related energy savings in one area (e.g., process control) being offset by ICT-related energy gains in another (e.g., increased freight transport). Additionally, their model incorporates elasticity values that temper energy savings potential with significant rebound effects [83].

Which future will manifest is hard to guess, with a recent review of macroeconomic studies showing inconclusive results [95, Table 1]. Rejeski [189] highlights ways in which ICT-enabled changes sweep beyond the effects usually analyzed in these studies, changing “the notions of property and ownership, the boundaries affecting jurisdiction, the dynamics of value creation, and the nature of competition.” The energy impacts of such systemic changes are all but impossible to quantify.

3.4 Conclusion

3.4.1 Persistent uncertainty in understanding the net energy effects of ICT

While both conceptual discussion and analytical modeling of ICT energy and environmental impacts have been occurring for nearly two decades, the jury is still out on the net effects of ICT adoption for several reasons. First, the complexity and variability of ICT deployment schemes make it difficult to isolate a standard implementation to analyze. Second, the lack of empirical data on how human users interact with ICT systems hinders the ability to assess actual energy effects. Third, the difficulties in disentangling the causes of interconnected effects lead to a tendency to fall back on theory—and on modeling exercises that conform to these theories, particularly where rebound is concerned. Finally, as the impact scope increases up the effect taxonomy (Table 3.1), the potential effect's magnitude and uncertainty increase dramatically. The emerging theme from service-specific studies suggests a consensus that ICT has large energy saving *potential*, but that the *realization* of that potential is by no means assured. In studies of rebound and systemic effects, the uncertainty only increases.

The variation of results in Tables 3.2-3.5 should drive home the conclusion that uncertainty plagues even the study of basic efficiency and substitution effects in fairly narrow, specific ICT applications. These differing results demonstrate a simple truth: it is possible to integrate ICT into a system in very inefficient ways—the mere addition of ICT to a system is not sufficient for net energy savings. The current state of understanding can be summarized with three related statements: the *technical potential* of ICT net energy savings is likely positive; the sign and magnitude of *realized* net energy savings are highly sensitive to the specific instantiation of ICT and how users interact with it; and, finally, the actual net energy effect is unclear and difficult to assess, especially when higher-order impacts are considered.

3.4.2 Research priorities

Though the overall net effect of ICT is likely to remain unknown, our review of the literature suggests several guidelines for improving the utility of research in this area, described below and summarized in Table 3.6. While some of these guidelines should already be normal research practice, they are not universally employed. Others will no doubt increase the burden on researchers and raise the bar for meaningful studies in this area; nonetheless, we believe their adoption is necessary to move the field towards greater understanding of ICT's true impact on energy use.

Collect and make publicly available data on energy use for a wide range of ICT technologies, strategies and systems. In a 2009 survey, the majority of experts rated the quality and availability of data to assess ICT's effect on energy efficiency as *Poor* or *Very Poor* [192]. Gathering more data in empirical studies allows assessment of how ICT systems are actually being deployed and used, further elucidating how specific conditions and parameters affect energy consumption and characterizing the "ICT energy savings gap"—the degree to which the realized energy performance of ICT fails to attain its estimated potential. At a broader level, a large-scale, survey-based data collection initiative similar to the Energy Information Administration's Residential Energy Consumption Survey (RECS) and Commercial Buildings Energy Consumption Survey (CBECS) for IT systems deployed in the residential, commercial, industrial, and transportation sectors would be helpful in providing insight into deployment strategies and baseline energy consumption measurements.

Use simulation and sensitivity analysis more broadly in impact estimates. Many studies use point values or relatively narrow ranges for input parameters. As a result, the estimated energy impacts reflect one or two specific views of ICT deployment and ultimately do little to advance the aggregate understanding of the ICT energy effect, since a different set of assumptions can usually be found that negates or reverses the findings. Exploring more of the solution space using stochastic modeling techniques such as Monte Carlo simulation would allow for statistically robust identification of driving factors and greater insight into the uncertainty surrounding such estimates. See Weber *et al.* [141] for an example of how LCA can be enriched with Monte Carlo techniques.

Pay more attention to study scope, particularly when comparing different studies. In lieu of providing a broader range of results in each study, one might argue that a sufficient set of separate point estimates can be aggregated to provide a bigger picture view. Such meta-analysis, indeed, formed part of the early vision for this paper. However, data sets, modelling methods, assumptions, and scopes vary so greatly as to make a straightforward synthesis of estimates nearly impossible. Thus, researchers should be diligent about exhaustively documenting their data, assumptions, and results so that others can replicate and adjust the results, if needed, for equitable comparison with other work.

Focus on identification of key parameters rather than aggregate impacts. Rather than focusing too heavily on estimating aggregate impacts—an exercise that, as this review shows, is unlikely to yield satisfying results—researchers should focus on identification of important parameters driving the energy use in ICT-infused systems (as several of the LCA studies do). Armed with such information, both public and private decision makers can design and implement intelligent, tailored, ICT-enabled systems that adapt to minimize energy use in deployment.

Increase focus on the behavioral aspect of ICT services. The studies here are generally technical in nature, depending heavily on assumptions about system structure and human behavior that may not reflect ground truth [116]. Focusing on behavioral aspects of ICT systems in concert with their technical properties would teach us how to align energy savings with user priorities. Amazon's shipping policies provide an illustrative example of how ICT pulls towards more energy consumption while also providing greater possibilities for reducing it through behavioral nudging. While faster delivery methods (e.g., same-day delivery, drones) are likely to be more energy-intense, Amazon has created incentive programs both for consolidating deliveries[9] and for using slower, more flexible freight modes where possible.[10] Since behavior can shift the direction of ICT impact, researching these sorts of ways to more precisely tailor ICT-enabled services to consumer needs could help temper the energy costs of ICT without appreciably sacrificing the quality of the customer experience.

Integrate higher-order effects. Few of the studies reviewed here address both second-order and third-order effects concurrently. Studies that present estimates of substitution and efficiency savings without addressing higher-order effects risk painting an overly simplistic picture of the ICT dynamic. Researchers should find ways to do more synthesis and integration across the taxonomy—i.e., evaluating possible higher-order effects in concert with an estimate of direct consumption, substitution, and efficiency. We can envision, for instance, a study examining whether the rapid growth of streaming video has increased the amount of content watched and placing this rebound estimate in the context of the direct energy use and substitution effects.

Notwithstanding these suggestions, developing an accurate and complete picture of the net energy effect of ICT remains a difficult task. However, we can continue to gain insight if we recognize that the specific implementation details, user behavior, and evolution over time are critically important and should not be oversimplified in the quest to compute an effect magnitude. Understanding the system dynamics as comprehensively as possible while remaining cognizant of limitations is a crucial step in ensuring that, as ICT continues its inevitable infusion in our economy and society, it functions as a dampener on energy consumption growth and a force multiplier for energy efficiency.

[9] "Subscribe-and-save" offers a discount on a periodic shipment (e.g., monthly) to replenish consumable goods. The predictability of the order allows Amazon to use slower modes as well as group recurring items into a single shipment. Amazon Pantry requires customers to fill a box with eligible items before it can be shipped, which incurs a flat fee. Add-on-items are small items that can be ordered through Amazon but that will not ship individually—they must be combined with other items. (Of course, some of these policies likely induce consumption, reducing the energy savings.)

[10] By offering video streaming credits and other incentives to customers who choose a "no-rush" delivery option.

Table 3.6: Guidelines for conducting future research on quantifying ICT indirect effects.

Issue	Remediation Guidelines
Lack of empirical data	• Conduct more empirical case studies; transition from "back-of-the-envelope" calculations and scoping studies. Focus on measuring realized savings rather than on estimating potential savings. • Broaden data collection and benchmarking programs such as DOE's energy consumption surveys and Center of Expertise programs [193] to collect and publish more comprehensive ICT deployment and energy use data. • In econometric work, focus on natural experiments to provide more evidence that results are ICT-driven. • Reconsider conducting studies where insufficient data to make robust conclusions exist. • Exhaustively document limitations and their anticipated effect on study conclusions.
Overly simplified point estimates	• Integrate Monte Carlo techniques to cover a broader range of inputs. • Conduct sensitivity analysis and report results. • Focus on identification of key parameters rather than on quantifying the aggregate impact.
Inconsistent system boundaries	• Refrain from face-value comparison of studies with different scopes. • When comparing new results to previous work, exhaustively document differences in data, methods, and assumptions. • Publish complete data, assumptions, and results so that others can fully replicate the study and can make adjustments to scope and assumptions to aid comparison with other work.
Narrow effect scoping	• Integrate higher-order impacts wherever possible. • Increase inclusion of behavioral aspects of ICT service deployment.

Chapter 4

Cost-Aware Load Shifting in Content Distribution Networks

> **Motivating question: to what extent can private and external electricity-related savings be achieved through real-time geographic load shifting in a network of data centers?**

When concerns are raised about the electricity consumption of a particular activity or facility, these concerns are typically related to the *cost* of the electricity. From the private perspective, electricity costs take the form of a power bill. From the public perspective, costs take the form of the impacts borne by society as a result of producing the electricity—i.e., the externalities associated with electricity generation. These externalities are the health, economic, and environmental impacts resulting from power plant emissions of criteria air pollutants [194] and greenhouse gases (GHGs), including human health consequences, reduced agricultural yields, reduced visibility, degradation of buildings and materials, and lost recreational value.

Since $cost = quantity * price$, there are two principal approaches to lowering these costs: reducing the amount of electricity used and reducing the price paid for the electricity. The previous chapter touched on the former approach: ways to reduce data centers' energy use. In this chapter, we examine the latter: reducing the private and external prices paid for electricity consumed in data centers.

The cost reduction strategy contemplated is arbitrage. Where their prices vary either temporally or spatially, goods can potentially be purchased at the lower price and sold at the higher price. Electricity exhibits this sort of variation: the U.S. electricity market is segmented regionally, and different regions of the country face different prices on the wholesale electricity markets. Additionally, because the grid must generally match generation to demand, wholesale market prices change hourly as the marginal generating unit on the dispatch curve—which is typically determined in merit (cost) order—changes between cheap baseload and more expensive load-following and peaking plants. This same dynamic leads to differences in the external costs of generation as well: different fuel mixes among the regions mean that some areas have different emissions intensities than others, while the emissions also change in time as different plants come online to meet demand.

However, it is not especially easy to transfer electricity over long distances from cheap to expensive areas due to transmission losses and grid constraints, nor is particularly cost-effective to store large amounts of electricity during times of low price for consumption when prices increase. (Hydroelectric plants are the obvious exception, though they are not an option likely to be available to data center operators.)

If the electricity itself cannot be easily shifted, what about shifting the *load*—the activity consuming the electricity? Traditional industries (manufacturing, aluminum smelting, etc.) can potentially shift load in time,[1] deferring production when electricity demand and prices are high, either in response to dynamic pricing or as part of a demand response program initiated by the grid operator. Internet data centers generally respond to real-time service requests and thus cannot shift their load temporally. However, data center networks can potentially do something physical industries cannot: shift load *geographically*. As noted by Armbrust *et al.* [57], "Physics tells us it's easier to ship photons than electrons; that is, it's cheaper to ship data over fiber optic cables than to ship electricity over high-voltage transmission lines."

Thus, we envision a geographically distributed network of data centers and analyze the cost savings available if the system operator were to shift computing load around the network in response to pricing signals. We compare private and external costs associated with different routing strategies—in particular, a strategy that minimizes private costs and a strategy that minimizes external costs.

4.1 Previous work

While there is a large volume of literature on how data center energy costs can be reduced through efficiency, this work falls within a smaller area of study on how the unique aspects of data center loads can be leveraged to reduce energy costs as well as offer services to electric utilities, particularly in shaving peak load and for ancillary services [195]. Shifting load among different data centers is by no means the only option in this regard. Liu *et al.* [196] note that data centers participating in coincident peak pricing

[1] Assuming that there is schedule slack: i.e., that maximum output of the plant running 24x7 exceeds demand.

demand response program can reduce utility load—and avoid peak pricing—simply by switching to the local backup generation maintained by these facilities, which is typically diesel. Thus, factoring the emissions impacts into the cost minimization is important from an environmental standpoint. Aksanli and Rosing [197] examine the ability of data centers to offer similar services using battery backup systems.

Ghatikar *et al.* [198] evaluate a range of seven data center demand response (DR) strategies, in which four real data centers are utilized as a demand-side resource for shifting a utility's peak load. However, the data centers in the evaluation are focused on storage and high-performance computing (HPC), and so are amenable to temporal load shifting in a way that Internet data centers are not. The bulk of the study is therefore focused deferring processing load to a time outside the DR window, though they also tested the scenario of shifting processing jobs from a data center participating in a DR event to a data center in a different area—a task conceptually similar to the load shifting strategy we examine here. However, saving the state of an HPC job, shifting it to a different location, and resuming processing is a use case with somewhat different constraints than shifting real-time loads.

In an overview of technologies needed for energy-efficient cloud computing, Berl *et al.* [199] mention load shifting as a key enabler for "energy-aware data centers." Rahman *et al.* [200] survey almost 30 studies related to geographic load balancing for data center power management, and Kong and Liu [201] identify 14 studies related to geographic load shifting in a review of "green" power management for data centers. Table 4.1 compares the studies most closely related with our work, each of which we summarize below. We omit from the table a variety of papers that limit themselves to development of conceptual algorithms for load shifting, but which do not meaningfully test these algorithms with real data [e.g., 202, 203]; these papers are useful for thinking about how private, external, and service costs may be integrated into an optimization model, but in this work we are focused on estimating the level of real-world savings that such algorithms could deliver. Therefore, studies that match real traffic patterns with actual cost and damage time series as model inputs are most applicable.

In what may be the most thorough treatment of load shifting in response to electricity price variability, Qureshi *et al.* [204] analyze the arbitrage opportunities created by variation in locational marginal prices (LMPs) in a network operated by Akamai, a leading content distribution network (CDN),[2] and find private cost savings ranging from 5% to 45% depending on assumptions regarding data center en-

[2] This type of network is described in more detail below.

Table 4.1: Comparison of this and other studies examining cost-aware load shifting.

Study	Traffic Data	Nodes	Elec. costs	External costs	Service costs
Qureshi *et al.* [204]	Akamai – 24 days / 39 months simulated (2008-2009)	9	ISO RTM LMPs	N/A	95/5 bandwidth constraint; parametric treatment of distances
Le *et al.* [205]	Ask.com – 1 day	3	Ameren RTPs; flat rates for solar & wind energy	Option to prioritize green energy	Enforcement of SLA on request service time
Zhang *et al.* [206]	Wikipedia – 2 months (2007); World Cup trace – 3 months (1998)	4	NYISO DAM LMPs	Wind and solar availability via meteorological data	None
Liu [207]	Hotmail – 2 days (2008)	14	Constant prices at each location	Zero marginal cost for renewables	Network and queuing delay
Gao *et al.* [208]	Akamai – 24 days	NR[a]	NR[b]	State-level average carbon emissions rate, with adjustment for daily peaks	95/5 bandwidth constraint; inclusion of distance (as latency proxy) in objective function
This paper	Akamai – 3 months / 1 year simulated	48	ISO RTM LMPs, translated into variable retail prices	Locational marginal & average damage estimates for CO_2, SO_2, NO_x, $PM_{2.5}$	Breakeven bandwidth costs; 95/5 constraint; parametric treatment of distances

[a] Not explicitly reported, but some subset of the same Akamai data used in Qureshi *et al.* [204].

[b] Electricity costs are included in their optimization model, but the source and type (average or marginal) are not reported.

ergy proportionality,[3] PUE, and bandwidth constraints. When peak bandwidth is constrained within existing levels, they find a maximum savings of 15%. Le *et al.* [205] find cost reductions of 25% under dynamic pricing while maintaining enforcement of a service level agreement (SLA) that a certain percentage of requests in each day must be met in a timely manner. They also investigate an option to prioritize green energy, finding it possible to reduce fossil energy use by about a third at only a 3% cost premium, but the assumptions and rough nature of their renewable energy scenario make it difficult to place much stock in this single data point.

Liu [207] finds private costs savings of 25-50% compared to a routing strategy that does not account for electricity costs, and he evaluates the possibility that computing load can “follow the renewables” by shifting load dynamically to locations with high wind and solar availability. The paper does not use dynamic pricing, however, incorporating geographic, but not temporal, price differentials. In the treatment of renewables, Liu assumes that data center operators own wind and solar plants and pay zero marginal

[3] For a discussion of energy proportionality and PUE, see Chapter 2.

cost for this generation, so substituting renewables for grid generation always reduces energy costs.

Zhang *et al.* [206] compare minimum-cost and maximum-renewable load shifting strategies against a “GreenWare” strategy that maximizes renewable generation while imposing a budget constraint. They assume that “green” energy is more costly than non-renewable “brown” energy, and therefore that increasing renewable energy use results in a direct increase in the power bill. The analysis finds that minimizing private costs reduces the power bill by roughly 1/3 to 1/2, while maximizing renewable energy usage increases the power bill by 30%. The GreenWare strategy successfully achieves about half of the potential increase in renewable energy use while preventing an increase in private costs. While this paper conceptualizes a means by which low-carbon electricity can be prioritized in traffic routing, the savings generated result from cost assumptions that are not necessarily valid. For example, the model inputs include electricity prices only from New York and apply these prices to four locations in Hawaii, California, Colorado, and Tennessee, an approach which is neither realistic nor likely to fully leverage differential costs, since prices within New York will be highly correlated with each other.

Gao *et al.* [208] go the furthest in modeling externalities by including carbon emissions as external costs in their optimization. They find potential carbon reductions of 5%-40% depending on PUE and bandwidth assumptions. The relationship between electricity prices and damages has important implications for the effects of load shifting. In regions where prices are positively correlated with damages, then minimizing one type of cost will reduce both. If, on the other hand, prices are negatively correlated with damages, the two cost-minimization strategies will tend to be at odds, meaning that efforts to reduce external costs will tend to increase the private costs incurred by the operator. While they use average prices and carbon emissions factors rather than marginal factors, Gao *et al.* [208] note that there appears to be little correlation between prices and emissions. They attempt to approximate temporal variation in emissions by assuming that peaking plants emit more carbon than base load plants, but this is not, in fact, universally the case.

Indeed, Holland and Mansur [209] find that measures which serve to flatten peak load, such as the real-time pricing (RTP) envisioned here, increase emissions in some regions of the U.S. This result occurs because, as we will see, marginal damage functions are not monotonically increasing with generation in the same way that marginal price curves typically are. That is, in regions where coal generation occurs earlier on the dispatch curve than natural gas, or where hydro is used to meet spiking demand, the marginal damage slope will be *negative*. In regions where high nuclear base load and renewables give way to increasingly dirty fossil generation—as Gao *et al.* [208] assume—the marginal damage slope will be positive.

The principal contribution of this work is an expansion of the cost optimization model to include a broader range of environmental impacts. While other papers have attempted to address external costs through allowing renewable generation to substitute for grid electricity or by using notional differentials in carbon emissions, this work is the only attempt, to our knowledge, to fully treat externalities in parallel with and in the same manner as electricity prices—that is, as a regionally differentiated, dollar-valued marginal cost linked to generation. Additionally, our external costs include those resulting not only from carbon dioxide but also from several criteria air pollutants. Finally, we use a larger, more comprehensive data set of real CDN traffic and electricity prices than previous work, we incorporate markups to translate wholesale prices into retail prices to represent more realistically the electricity costs incurred by the data center operator, and we use the latest electricity price and damage data available.

4.2 Methods, model definitions, and metrics

In this section, we define a series of routing models that will allow us to assess energy savings opportunities for load shifting, as well as establish the methods and metrics used to assess model performance.

4.2.1 General problem statement

Perhaps the best application for the load shifting program envisioned here is a content distribution network (CDN). A CDN is a large, highly-distributed network of servers that delivers web traffic (e.g., web pages and streaming video) from content providers (e.g., Google, Amazon, and Youtube) to consumer-facing Internet Service Providers (e.g., Verizon and Comcast) and ultimately to consumers. CDNs typically speed delivery through "edge-caching": pushing replicated content closer to the "edge" of the network, reducing the number of "hops" along the traffic's path and avoiding bottlenecks [210]. CDNs are quickly becoming the dominant delivery method: the proportion of all Internet traffic carried by CDNs will grow from 45% in 2015 to 64% by 2020 [211]. For this reason, electricity consumption by CDNs is likely growing faster than the low growth rate seen by data centers as a whole.

We start with a CDN consisting of a set of data centers (nodes) and a set of clients with variable demand for web traffic that must be met by the network nodes and then explore the costs associated with different strategies for routing that traffic. While the routing strategies are different, the central method used in this analysis is optimization, with the following general problem statement:

> Minimize the targeted operating cost component of the CDN by selecting the volume of traffic to send from each node to each client, subject to node, bandwidth, and distance constraints as required.

The diagram in Figure 4.1 shows the different components of this optimization model, which we now describe briefly.

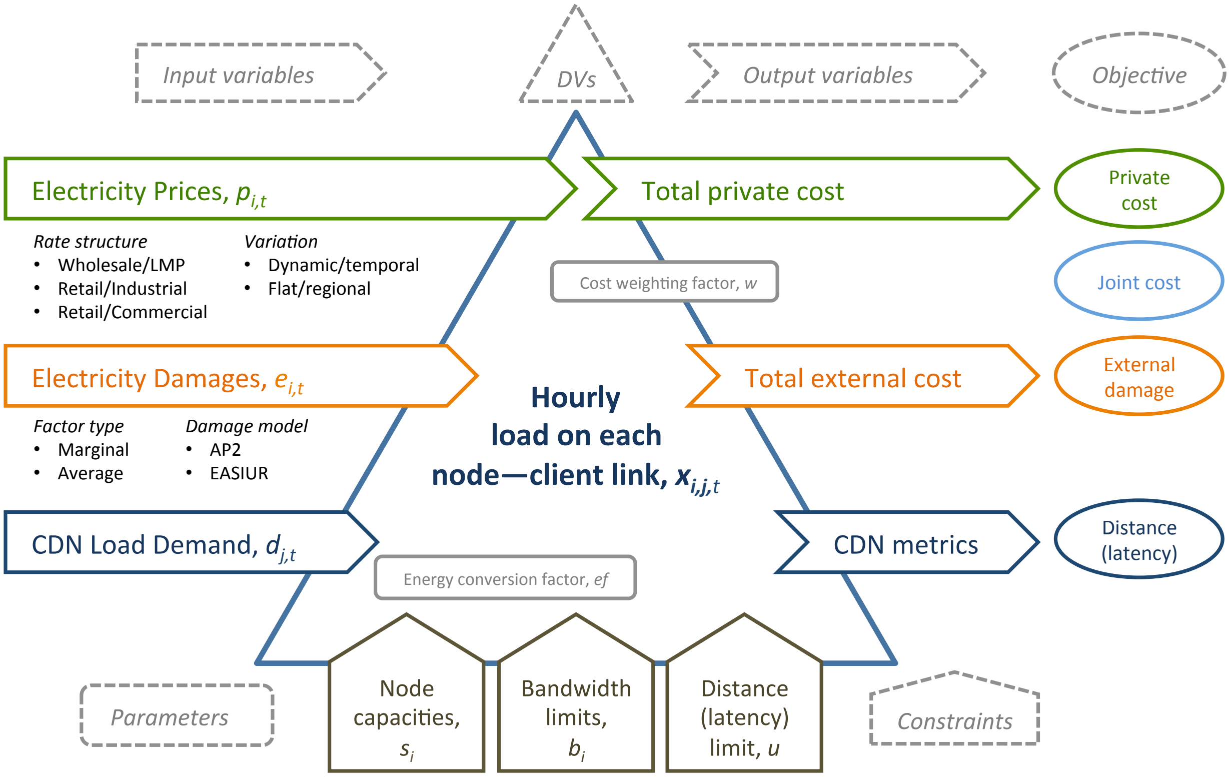


Figure 4.1: Overview of model components. Subscripts i, j, and t correspond to node, client, and hour indices, respectively.

Decision variables (DVs). At the center of the model is a cost-minimizing decision process that chooses the CDN traffic routing—that is, how much traffic to send over each node-client link in each hour.

Input variables. The minimizer takes as inputs three separate hourly time series: traffic demand at each client, electricity prices at each node, and electricity damages at each node. We discuss each of the inputs in greater detail in Section 4.3, and thus only briefly summarize them here. The traffic input comes from Akamai's CDN and is converted to energy using an energy conversion parameter, *ef*. Each of the two cost metrics have several different versions. We use three different rate structures for electricity prices: a LMP that may be thought of loosely as the wholesale price, the industrial retail rate, and the commercial retail rate. Each of these rates, in turn, may be either flat, varying only by region and remaining relatively constant over the course of the year, or dynamic, varying both by region and hourly. The damage input can reflect either marginal or average damages estimated using either of two different damage models, AP2 and EASIUR.[4]

Output variables. Outputs of the optimizer include not only the routing solution, but also the private and external costs of electricity, and various CDN metrics such as overall transport distance (as a proxy for latency) and peak bandwidth used.

Objectives. The program may be used to minimize only the private cost of electricity, only the external cost, or any joint weighting of the two by adjusting the weighting parameter, w. Additionally, the program can optimize a CDN performance metric, such as distance. No matter which objective is selected, all costs and metrics are calculated for each solution, so that, e.g., the effect of minimizing external costs on private costs can be observed.

Constraints. The set of feasible solutions to the problem may be constrained by maximum capacities at each node, bandwidth limits, and distance limits.

We run a suite of scenarios that employ different permutations of this general model, with different combinations of outputs, cost metrics, and constraints. Each set of scenarios is described briefly below and summarized in Table 4.2.

4.2.2 Baseline strategies

Before defining the load shifting strategies, we first identify baseline strategies to benchmark energy costs against which the cost-minimization strategies will be compared.

[4] Air emissions models must require contrived acronymic names to get published: AP2 is version 2 of APEEP, which stands for *Air Pollution Emission Experiments and Policy analysis*; EASIUR stands for *Estimating Air pollution Social Impact Using Regression.* These models are described in more detail in the data section.

Actual routing

The primary baseline strategy is simply the actual routing used in the Akamai traffic data. This loading solution is the output of Akamai's internal routing algorithm, which is black box for this analysis but generally factors in latency (proximity), packet loss, and cost of service (primarily bandwidth).

The energy costs of this model are obtained simply by converting traffic at each node into energy consumption and applying the appropriate cost of that energy—either private, external, or social (both). No optimization is done on our part.

Proximal routing

We may wish to approximate the CDN routing algorithm on any generic traffic load—that is, a load for which we don't know the CDN's actual routing. Such a rule-based strategy would allow creation of a realistic "baseline" routing strategy for any arbitrary traffic load, rather than being limited to actual traffic data provided by a CDN. CDNs primarily use edge caching, meaning that—all else being equal—content will generally be served from the nearest node. Of course, other factors such as capacity constraints, bottlenecks, minimization of packet loss, and load balancing mean that actual routing is not this simple. We evaluate how well a proximity-based strategy matches the CDN's actual routing. Referring to Figure 4.1, the model minimizes distance in the objective function subject to node capacities.

Formally, the proximal routing strategy is a minimization of the total network transport in GB-mi:

minimize:

$$f(\mathbf{X}) = \sum_{i,j,t} \left(l_{i,j} * x_{i,j,t} \right) \tag{4.1}$$

with respect to $x_{i,j,t}$ and subject to:

$$\sum_{i} \left(x_{i,j,t} \right) = d_{j,t} \qquad \forall j,t \tag{4.2}$$

$$\sum_{j} \left(x_{i,j,t} \right) \leq s_i \qquad \forall i,t \tag{4.3}$$

$$\mathbf{X} \geq 0 \tag{4.4}$$

where:

$$\begin{aligned} x_{i,j,t} &= \text{ the traffic from node } i \text{ to client} j \text{ in hour } t \\ l_{i,j} &= \text{ the distance between node } \ i \text{ and client } j \\ d_t &= \text{ the service demand at time } t \\ s_i &= \text{ the service capacity at location } i \end{aligned}$$

4.2.3 Energy cost minimization strategies

We now define several load shifting strategies as cost-minimization problems. The strategies are formulated as linear programs (LPs) with the general form:

minimize:

$$f(\mathbf{X}) = \sum_{i,t} \left(\left(w * p_{i,t} + (1-w) * e_{i,t} \right) * ef * \sum_{j} \left(x_{i,j,t} \right) \right) \tag{4.5}$$

with respect to $x_{i,j,t}$ *and subject to:*

$$\sum_{i} \left(x_{i,j,t} \right) = d_{j,t} \qquad \forall j,t \tag{4.6}$$

$$\sum_{j} \left(x_{i,j,t} \right) \leq s_i \qquad \forall i,t \tag{4.7}$$

$$\mathbf{X} \geq 0 \tag{4.8}$$

where:

$$\begin{aligned} x_{i,j,t} &= \text{ the traffic from node } i \text{ to client} j \text{ in hour } t \\ p_{i,t} &= \text{ the private per-unit cost of electricity in location } i \text{ at time } t \\ e_{i,t} &= \text{ the external per-unit cost of electricity in location } i \text{ at time } t \\ d_t &= \text{ the service demand at time } t \\ s_i &= \text{ the service capacity at location } i \\ ef &= \text{ the conversion factor, energy consumption per unit of traffic volume} \\ w &= \text{ a factor determining how private and external costs are weighted in the objective} \end{aligned}$$

We minimize different electricity costs by altering the value of w in this model, and we perform multiple runs with the different cost input data shown in Figure 4.1 and described in Table 4.5. This problem coded into GAMS and solved using the CPLEX linear solver. In all cases, the optimization code is written such that private and external costs (p and e) are separately calculated and reported as outputs.

Minimizing the private energy cost

The private cost minimization strategy focuses on reducing the CDN operator's power bill. This strategy is formally modeled by setting $w = 1$ in Equation 4.5, in which case only private costs are considered in the objective function. The node capacity constraint is active; bandwidth and distance are unconstrained. We evaluate the cost savings under the six different electricity pricing schedules listed in Figure 4.1 and described in more detail in Section 4.3 and Table 4.5.

Minimizing the external energy cost

The external cost minimization strategy reduces damages associated with the data centers' electricity consumption. In this case, $w \approx 0$ in Equation 4.5; only external costs are considered. Because external costs are estimated regionally (see Section 4.3.3), states in the same region face the same external cost profile. In order to break such ties, we set w to a very small positive value so that, in cases where two data centers have the same damage value, the location with the least expensive electricity is preferred. There may still be some ties in cases where the model is using EIA retail prices, which are also regional values; however, in such cases it does not matter which data center is selected since all costs are equivalent.

The node capacity constraint is active; bandwidth and distance are unconstrained. The model is run minimizing each of the four different versions of electricity damages.

Minimizing the social energy cost

Finally, social costs can be minimized by setting $w = 0.5$, in which case the program minimizes the sum of private and external costs. The relative weights of private and external costs can be altered by choosing other values of w. As above, the node capacity constraint is active; bandwidth and distance are unconstrained.

4.2.4 Accounting for other costs

The key performance metrics of this analysis are private and external electricity costs; strategies with lower costs are preferred to those with higher costs. However, minimizing electricity costs can affect other costs associated with load shifting, and so we run special cases of the strategies outlined above with additional constraints to explore the impact of these other considerations, focusing on latency and bandwidth.

Latency

Minimizing electricity costs will likely increase the mean client-server distance, which will in turn increase the service latency—the time it takes for the data to reach the client. Latency increases with distance for several reasons, including propagation delay, which is simply determined by the signal transmission rate in the medium (e.g., fiber optic cable), and delays caused by amplification and switching—relaying the signal across mutliple network "hops." Latency is measured in round-trip time (RTT), or the time it takes for an information packet to be transmitted from one network node to another and an acknowledgement to be sent back. Decreasing latency is a primary reason for the existence of CDNs and why they use edge-caching.

Since increased latency may affect customers' value of the service and thus CDN revenue, the CDN operator may wish to limit the geographic range that can be served by each data center. To investigate a latency-constrained scenario, we add the following distance constraint to the optimization model described in Equation 4.5:

$$x_{i,j,t} = 0 \qquad \text{where } l_{i,j} > u \tag{4.9}$$

where:

$$\begin{aligned} l_{i,j} &= \text{ the distance between node } i \text{ and client } j \\ u &= \text{ a constant denoting the distance limit for continental U.S. traffic} \end{aligned}$$

This constraint ensures that no traffic is loaded onto any link with a node-client distance greater than u.

Bandwidth

One type of web traffic that is tolerant of high latency is streaming video, because it can be buffered, and so it would seem that video is a particularly good use case for load shifting. CDNs are the preferred method for delivering video content, carrying 61% of it in 2015 and expected to handle 73% of it in 2020 [211]. Unfortunately, video faces another issue: bandwidth. Cisco estimates that Internet video was 70% of traffic in 2014 and will increase to 82% by 2020 [211]. Sandvine similarly reports that 70% of North American Internet peak evening traffic is streaming video, up from 35% five years ago [212]. This volume is high enough that shifting it back and forth over the long-haul links would likely cause bottlenecks, which is precisely the reason edge-caching is so valuable and why metro traffic growth is outpacing that of long-haul traffic [213]. In fact, Netflix's Open Connect Appliances replicate of large portions of the Netflix library at data centers owned by internet service providers (ISPs), close to end-users, to avoid these bottlenecks [214, 215]. For such traffic, a distance constraint, as in Equation 4.9, could be employed to keep traffic near the edge for bandwidth as well as latency reasons.

However, general web traffic can likely be shifted. We know that shifting will increase the aggregate transmission distances and that this increase may have an effect on revenue from certain types of traffic, as considered in the previous section. However, such an increase in distance does not affect how much the CDN would pay for transmission, as these charges are insensitive to distance. Rather, CDNs pay for bandwidth, and the typical billing mechanism is a peak 95/5 structure where they pay for the 95th percentile of bandwidth used.[5] In other words, the colocation provider samples the traffic rate over the course of a month, throws out the highest 5% of the samples, and charges for the highest remaining traffic rate.

The implication of this billing model for load shifting is important: any node that has a favorable energy cost and is utilized by the load shifting optimizer more than 5% of the time will face the maximum possible bandwidth bill, and the more dynamic the solution is, the more nodes in the network will fall under this scenario. This fact might make load shifting much less appealing. In addition to evaluating the impact of cost minimization on the bandwidth percentile at each node, we add a constraint to the optimization model to limit the increase in peak bandwidth to the baseline 95th percentile. This version of the model will tell us the level of electricity cost savings that can be achieved without increasing bandwidth costs at all.

However, because the bandwidth constraint is percentile-based, making this adjustment is not as straightforward as was adding the distance constraint (Equation 4.9). That is, the bandwidth constraint is a soft constraint that may be violated 5% of the time. This class of optimization problem is known as a *k-violation linear program*, which is generally solved using mixed-integer programming (MIP) and is NP-complete, although there are improvements that can be made for certain subclasses [216].

We supplement the model by adding the following constraints, variables, and constants:

[5] Unless otherwise noted, *peak* bandwidth implies the 95th percentile of bandwidth usage.

$$\sum_j \left(x_{i,j,t}\right) < b_i + c_{t,i} * s_i \qquad \forall i, t \tag{4.10}$$

$$\sum_t \left(c_{t,i}\right) < k_i \qquad \forall i \tag{4.11}$$

where:

$b_i =$ the 95$^{\text{th}}$ percentile bandwidth constraint for node i
$c_{t,i} =$ a binary variable indicating whether or not b_i is broken in hour t at node i
$k_i =$ a constant denoting the number of hours in which node i may exceed b_i

As before, $x_{i,j,t}$ is the traffic load served from node i to client j in time t, and s_i is the capacity of node i. The LHS of Equation 4.10 is thus the total traffic served out of each node in each hour.

These two additional constraints represent a "big-M" approach for selectively enforcing constraints. Equation 4.10 is the bandwidth constraint. The limit b_i is set to the 95$^{\text{th}}$-percentile bandwidth observed at each node in the baseline case. As noted, however, b_i is not a hard constraint. The binary variable $c_{t,i}$ is "on" (equal to 1) when the limit b_i is exceeded. When $c_{t,i} = 1$, the bandwidth limit is increased by s_i—the "big M." Mixed integer solvers benefit by reducing the potential solution space, so we adopt the best practice of setting M only as large as it needs to be; in this case, traffic cannot exceed the capacity of the data center, s_i, so we set $M = s_i$ for each node. Equation 4.11 restricts the number of times b_i can be exceeded at each node to no more than k_i, where k_i is set to 5% of the number of hours in the time series. Between these two equations, then, the model enforces the 95$^{\text{th}}$ bandwidth constraint.

The additional complexity of the MIP means that it takes much longer to solve than do the LP versions of the model. The main reason for the added complexity is addition of $c_{t,i}$ to the problem. In the LP version, each hour of the year is independent. In the MIP formulation, we have now linked the bandwidth constraints, Equation 4.10, at each node through Equation 4.11. The optimizer must now select, for each node, in which k hours of the time series to violate the bandwidth limit, which is essentially a knapsack problem with the hour subscript, t, driving the combinatorial explosion.

For this reason, we solve separately for each of the twelve months to make the problem more tractable. The 95/5 bandwidth is typically calculated and billed monthly, so this modification does not result in the need for any additional assumptions or abstractions in the model. We sum the monthly electricity bills under the bandwidth-constrained model and compare its savings with those of the unconstrained LP model. To keep the runtime manageable, we use a gap tolerance of 1.5% for the MIP solver, which means that the solutions are guaranteed to within than 1.5% of optimal.[6] Thus, the savings reported under the bandwidth-constrained model is a lower bound—actual savings may, in fact, be up to 1.5% greater than indicated.

4.2.5 Summary of scenarios

The scenarios outlined above can be grouped into three main categories:

1. Baseline
2. Electricity cost minimization
3. Constrained

Category 1 contains baseline strategies—no cost minimization is used beyond Akamai's existing algorithm. Category 2 evaluates the maximum potential of load shifting to minimize electricity costs. Category 3 contains special cases of Category 2, in which the load shifting optimizer is further constrained by other routing considerations. Note that the capacity constraints of each node are implemented in *all* scenarios, while Category 3 imposes additional constraints. Table 4.2 provides a quick reference on which scenarios implement which parts of the general model depicted in Figure 4.1, and the next section describes the input data used for these scenarios.

4.3 Data sources

As depicted in Figure 4.1, the model relies on three main input data sets: (1) a data center network topology and network traffic load, (2) an hourly time series of electricity prices, and (3) a corresponding hourly time series of electricity damages. In general, the prices and damages are *marginal*, that is, they reflect the cost associated with consuming an additional unit of electricity in that particular hour; however, we also explore the use of marginal vs. average damage factors. We aggregate data at the state level; our final data set has hourly prices and damages for each state, and we assume that each state has a single data center node.

[6] This threshold seemed to be a good tradeoff between accuracy and runtime. The solver can generally find a solution for a one-month time series within 1.5% in on the order of 15-30 minutes, while searching for a solution within 1% took as long as 5 hours during trial runs. For this application, the payoff for the additional accuracy is not particularly high—putting an upper bound estimate on cost is good enough, given uncertainty elsewhere in the problem.

Table 4.2: Summary of model scenarios. Constraints: *Cap.* = node capacity, *BW* = bandwidth, *Dist* = distance. Electricity prices: *Dyn.* = dynamic, *L* = LMP, *I* = industrial retail, *C* = commercial retail. Damages: *Avg.* = average, *Mar.* = marginal, *A* = AP2 model, *E* = EASIUR model.

Scenario	**Min. Objective**	**Constraints**			**Electricity Price**						**Damages**			
					Flat			*Dyn.*			*Avg.*		*Mar.*	
		Cap.	BW	Dist	L	I	C	L	I	C	A	E	A	E
Category 1: Baseline														
Actual	N/A	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Proximal	Transport (GB-mi)	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Category 2: Electricity cost minimization														
Private cost minimization	Private costs	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
External cost minimization	External costs	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Joint minimization	Weighted costs	✓							✓					✓
Category 3: Constrained														
Constrained distance	Private & external costs	✓		✓					✓					✓
Constrained bandwidth	Private costs	✓	✓						✓					✓

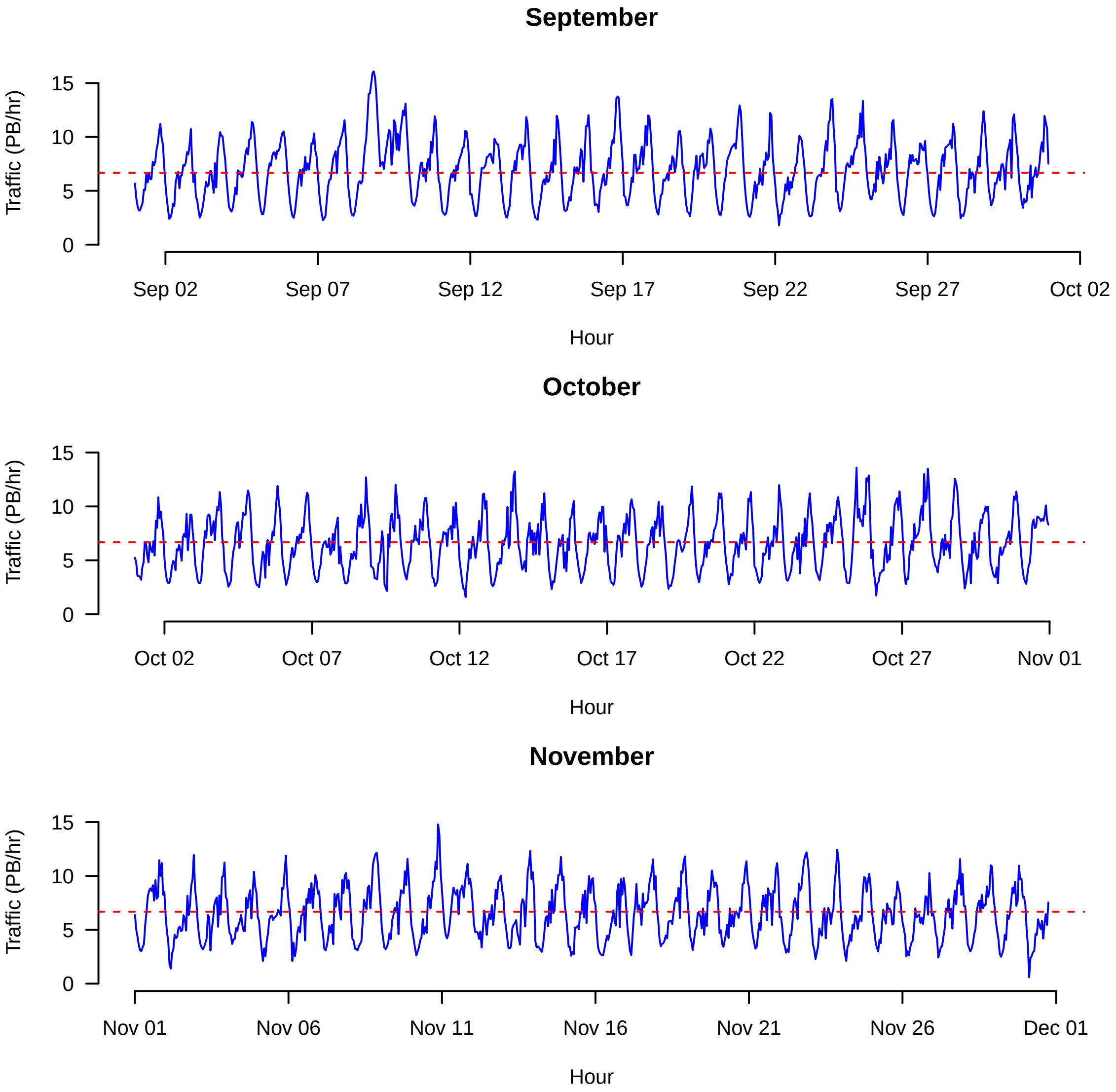
4.3.1 Network topology and traffic data

We obtained three months of traffic data from Akamai, a leading CDN. The data contains hourly traffic volume served from September to November of 2015 by 906 data center nodes, called ECORs, in 333 U.S. cities to 294 client locations around the world. ECOR stands for Equivalence Class of Regions and represents a group of proximal server racks (called "regions") that serve similar loads. We will call each ECOR or data center a *node* and each aggregated client location a *client*. Each node has a capacity, which can change over time due to addition or removal of servers, maintenance downtime, or unplanned outages, and each client has an hourly demand.

Before examining the distribution of traffic among nodes and clients, it is instructive to examine the overall traffic pattern, shown in Figure 4.2. We can observe a strong diurnal load pattern and indication of unusually high load from September 8-9. Exploratory data analysis revealed further revealed a small weekly pattern, with traffic being a bit higher on Tuesday, Wednesday, Saturday, and Sunday, than on Monday, Thursday, or Friday.

We aggregate node capacity at the state level (excluding Alaska and Hawaii) and client demand at the level of a U.S. state, Canadian province, or foreign country, resulting in an aggregated data set of 48 nodes and 69 clients. The top twenty clients by traffic volume are shown in Table 4.3.

Our initial analysis uncovered an error in the node capacity data field. Therefore, we set the capacity of each data center at its highest observed traffic volume; this may be a conservative estimate, as there



September

October

November

Figure 4.2: Three-month hourly profile of Akamai web traffic at U.S. data centers. Data from 2015; shaded blocks indicate weekend days; dashed line is mean traffic level. Traffic level has been scaled up to the order of magnitude matching Akamai's total U.S. web traffic.

Table 4.3: Share of web traffic by client location (top 20). All other client locations (49) each constitute 1% or less of traffic.

Rank	Client	Share of traffic
1	California	11%
2	South America	10%
3	Texas	8%
4	New York	6%
5	Florida	5%
6-7	Asia, Illinois	4% (each)
8-13	Virginia, New Jersey, Georgia, Pennsylvania, Washington, Other North America	3% (each)
14-20	Ohio, North Carolina, Michigan, Massachusetts, Arizona, Colorado, Maryland	2% (each)
	Total	78%

may in fact be additional reserve margin at each node. Figure 4.3 shows a histogram of data center capacity by state. The top nine nodes labeled in the figure collectively account for 86% of total traffic capacity.

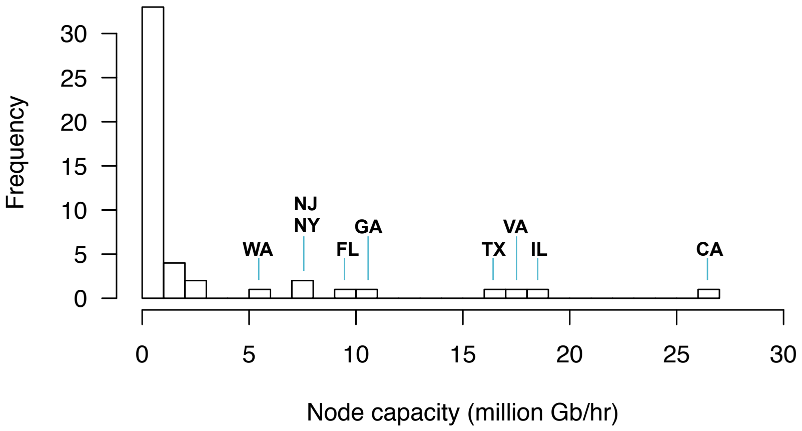


Figure 4.3: Histogram of node capacity by state.

Figure 4.4 shows the distribution of hourly utilization rates for the pooled state-aggregated nodes. The mean and median utilization rates are approximately 30%, which is in line with published estimates of data center utilization.

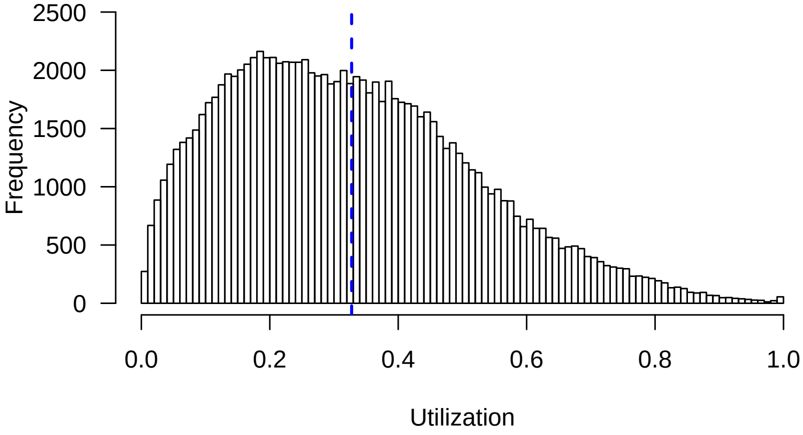


Figure 4.4: Distribution of state-level node hourly utilization rates. The blue dashed line is the mean utilization rate of 30%.

Traffic simulation

Our model requires a year's worth of data to fully capture seasonal variation in electricity prices and damages. We sample from the three-month time series to create a full year of traffic. To preserve both the daily and weekly trends, we sample by day and hour. For example, to simulate traffic load by the Pennsylvania client on any Wednesday at noon, we draw a random observation from the population of Wednesday Pennsylvania traffic loads from 12 – 1 p.m. in the three-month data set. After the simulation, we have a year's worth of hourly traffic loads for each of the 69 clients.

The overall traffic volume in the data set is several times less than Akamai's overall U.S. traffic load. Therefore, we scale up the load to the order of magnitude of Akamai's U.S. traffic of 60 million TB.[7] The most important aspect of the data is the hourly traffic pattern, so we do not anticipate that this scaling will materially affect the results—it will simply determine the absolute size of the overall energy cost, which will scale linearly with this input. Therefore, costs for lower or higher traffic volumes can be easily estimated by rescaling the results.

[7] A rough estimate provided by Bruce Maggs of Akamai.

Converting traffic load to energy consumption

This traffic load will be allocated to data center nodes using optimization strategies that minimize energy costs; therefore, at each node, the traffic must be converted into energy consumption. LIke Gao *et al.* [208], we assume that, while individual servers are not likely energy proportional, data centers approximate energy proportionality at the facility level. Recent estimates of energy use per unit of traffic served at the data center range from 1 kWh/GB [217] to 2.5 kWh/GB [218], with earlier estimates of the energy factor being even higher. However, back-of-the-envelope calculations as well as conversations with data center operators indicate that these estimates are almost certainly too high for a CDN. In their supplemental information, Malmodin *et al.* [217] note that the energy factor for an efficient, video-streaming data center might be drastically lower, at 0.01 kWh/GB. Hunt [219] claims that Netflix's energy consumption per streaming hour was 0.0013 KWh in 2014, and a 1 GB/hr streaming rate is not an unreasonable estimate for the data rate of standard definition video. If we assume 3 GB/hr for high-definition video, then the energy consumption is 0.004 kWh/GB.

Thus, the range of conceivable energy conversion factors varies over several orders of magnitude, a finding echoed in Aslan *et al.* [21], which evaluates energy conversion factors for data transmission over networks. Somewhat arbitrarily, we use an energy factor of 0.01 kWh/GB, under the assumption that Akamai's IT infrastructure is very energy efficient. This should result in a fairly conservative estimate of energy costs, as they could be much higher if the conversion rate is larger. However, the energy cost results will scale linearly with this factor, so the effect of different values on overall costs can be easily estimated.

4.3.2 Price data

The U.S. electricity grid is broken up into balancing authorities (BAs), which historically comprised the service areas of vertically-integrated utilities. Each BA is responsible for maintaining the grid's balance (i.e., supply = demand) in its service area, which can range from small, covering a single city, to very large, spanning multiple states. There are two fundamental models for electricity markets in the U.S. Utilities in regulated markets typically remain vertically integrated and charge retail electricity prices that are approved at the state level by a Public Utilities Commission (PUC), and which cover the cost of service plus a rate of return for the utility. In restructured markets created by industry deregulation beginning mid-1990s, utilities are generally "unbundled," with the legacy utility retaining the distribution, customer billing, and provider-of-last-resort functions, while generation is opened up to a competitive marketplace. Customers have "electricity choice" and can select their generation provider based on cost, environmental performance, or other factors.

Each of these regional marketplaces is managed by a regional transmission organization (RTO) or independent system operator (ISO),[8] which operates, at a minimum, two wholesale electricity markets: a day-ahead market to allocate generation for the next day, and a real-time "balancing" market to meet unmet demand as needed. (Some ISOs operate additional markets at different frequencies—e.g., a fifteen-minute market as well as an hourly market—as well as other types of markets for forward capacity and ancillary services such as reserves and regulation.) The goal of the ISO is to create an efficient, transparent market for electricity. Thus, generators bid into the market and are dispatched in economic or "merit" order of increasing cost, with all generators paid at the clearing price [220]. Restructuring has been a major driver behind the dissolution and consolidation of smaller BAs into the regional wholesale organizations; the number of BAs has declined from over 120 in 2006 to around 60 in 2014 [221]. The regions of the country covered by ISOs are the shaded regions in Figure 4.5; the remaining areas are regulated.

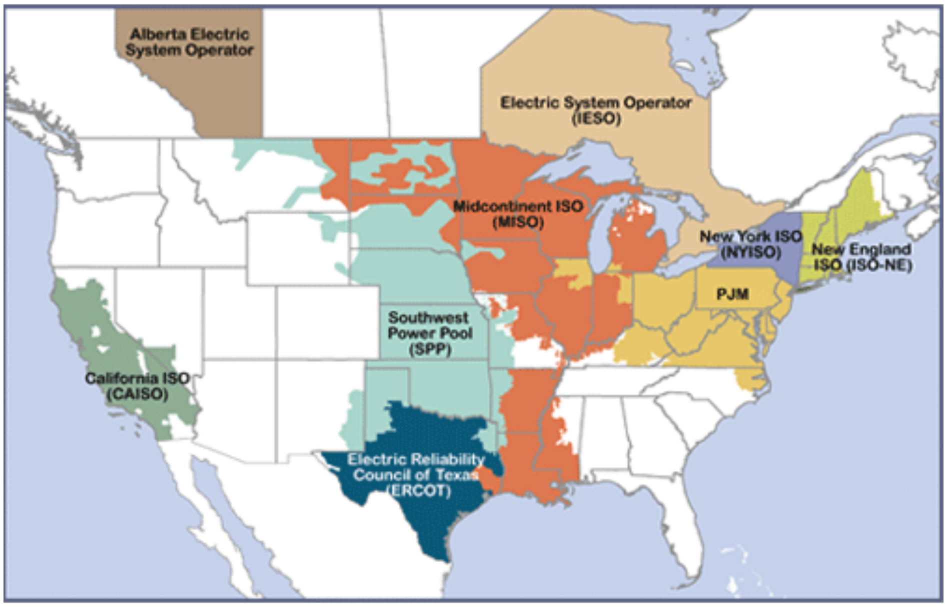

Alberta Electric System Operator
Electric System Operator (IESO)
Midcontinent ISO (MISO)
New York ISO (NYISO)
New England ISO (ISO-NE)
PJM
Southwest Power Pool (SPP)
California ISO (CAISO)
Electric Reliability Council of Texas (ERCOT)

Figure 4.5: U.S. ISO/RTO region map. Source: FERC / US Government work / Public domain.

[8] The difference between an RTO and an ISO is somewhat subtle and not relevant for the purposes of this work. We will generally use the term ISO to encompass both.

It is useful to introduce two other electricity grid maps for comparison. Figure 4.6 (top) shows the continental U.S. broken into eight regional reliability organizations under the auspices of the North American Electric Reliability Corporation (NERC), the organization tasked with maintaining grid reliability. Some of the ISO footprints are coterminous with the NERC regions, but there are large boundary differences in other areas—particularly MISO. The NERC regions can be further subdivided into subregions, shown in the bottom of Figure 4.6, which are used by the EPA's Emissions & Generation Resource Integrated Database (eGRID) as well as some of the EIA's market price reporting.

Locational marginal prices

At the heart of the restructured markets are electricity prices, which are called Locational Marginal Prices (LMPs) because they are:

- *Locational*: prices vary among nodes, which represent specific locations in each market. Different types of nodes include generation nodes, which are prices at generators; buses, at transmission points; load zones, or groups of electricity consumers; hubs or aggregate pricing nodes (APNs), which are aggregated, weighted prices for a particular region; and interfaces, prices in neighboring regions at which electricity can be imported.
- *Marginal*: prices reflect the cost of procuring an additional unit of electricity at that point in time.

LMPs are typically composites of three price components. The energy component reflects the cost of generation, the congestion component accounts for additional costs imposed by grid capacity constraints, and the loss component includes line losses. Thus, grid congestion and distance from generators will tend to increase a node's LMP. As noted above, the energy component of the LMP is the clearing price in the energy market. Generators' bids will cover variable costs (and are thus highly correlated with fuel costs), startup and shutdown costs, recovery of fixed costs, and return on investment [220].

Historical LMPs are archived by the ISOs and generally available online, and we use the LMPs, averaged within each hour, from the real-time markets at the primary hub or APN in each state. Thus, it is fairly easy to obtain LMPs for the portion of the country managed by ISOs (Figure 4.5).[9] Regions of the U.S. that do not fall under an ISO are typically governed by bilateral contracts between generators and

[9] Although the different ISOs' interfaces for obtaining the LMPs have varying degrees of ease; PJM, IS-ONE, and NYISO are painless, while SPP and—surprisingly—CAISO, are quite frustrating.

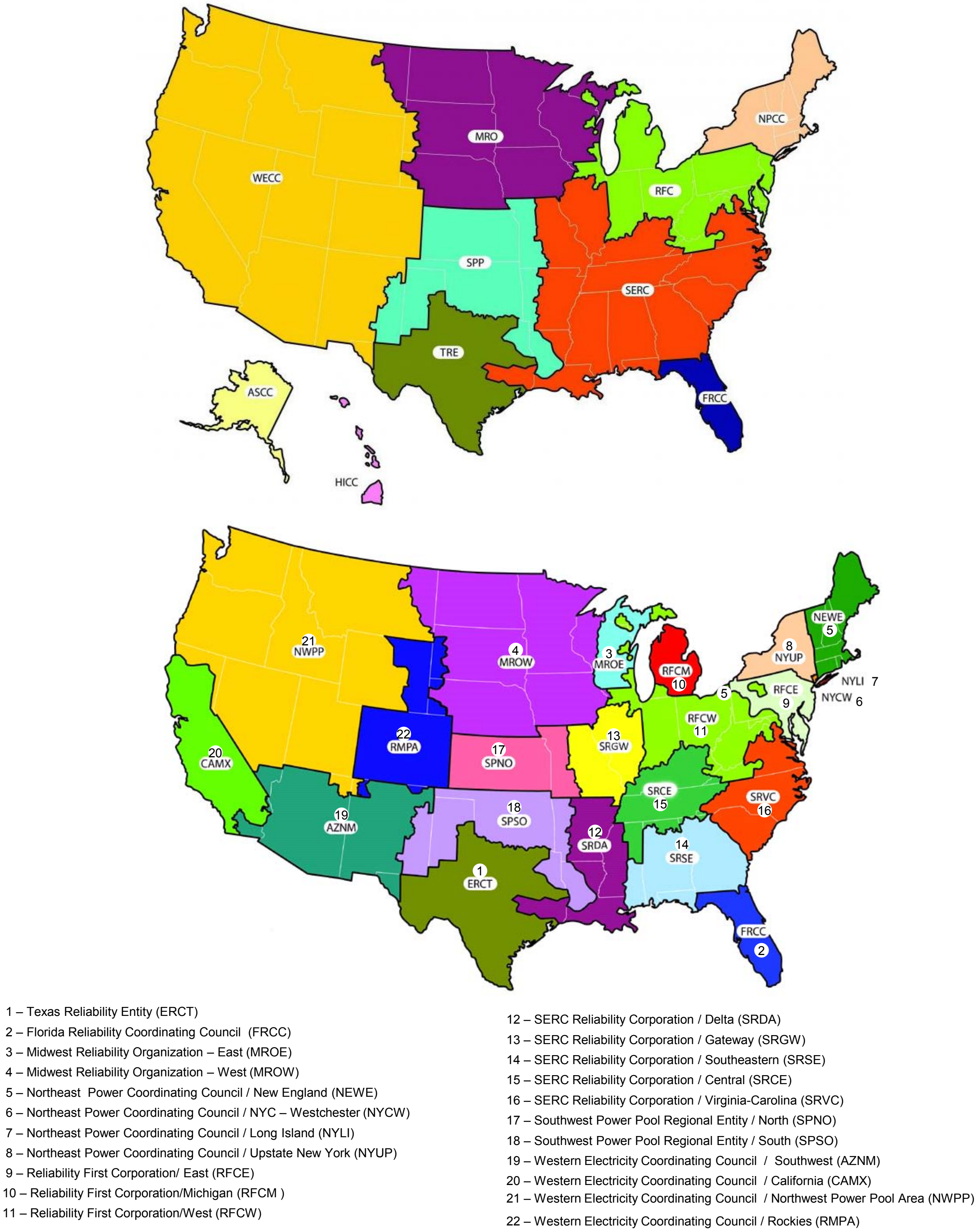

WECC
MRO
NPCC
RFC
SPP
SERC
TRE
FRCC
ASCC
HICC
21 NWPP
22 RMPA
20 CAMX
19 AZNM
4 MROW
3 MROE
RFCM 10
5
NEWE 5
8 NYUP
RFCE 9
NYLI 7
NYCW 6
RFCW 11
13 SRGW
17 SPNO
18 SPSO
12 SRDA
1 ERCT
SRCE 15
SRVC 16
14 SRSE
FRCC 2
1 – Texas Reliability Entity (ERCT)
2 – Florida Reliability Coordinating Council (FRCC)
3 – Midwest Reliability Organization – East (MROE)
4 – Midwest Reliability Organization – West (MROW)
5 – Northeast Power Coordinating Council / New England (NEWE)
6 – Northeast Power Coordinating Council / NYC – Westchester (NYCW)
7 – Northeast Power Coordinating Council / Long Island (NYLI)
8 – Northeast Power Coordinating Council / Upstate New York (NYUP)
9 – Reliability First Corporation/ East (RFCE)
10 – Reliability First Corporation/Michigan (RFCM)
11 – Reliability First Corporation/West (RFCW)
12 – SERC Reliability Corporation / Delta (SRDA)
13 – SERC Reliability Corporation / Gateway (SRGW)
14 – SERC Reliability Corporation / Southeastern (SRSE)
15 – SERC Reliability Corporation / Central (SRCE)
16 – SERC Reliability Corporation / Virginia-Carolina (SRVC)
17 – Southwest Power Pool Regional Entity / North (SPNO)
18 – Southwest Power Pool Regional Entity / South (SPSO)
19 – Western Electricity Coordinating Council / Southwest (AZNM)
20 – Western Electricity Coordinating Council / California (CAMX)
21 – Western Electricity Coordinating Council / Northwest Power Pool Area (NWPP)
22 – Western Electricity Coordinating Council / Rockies (RMPA)

Figure 4.6: NERC entity region (top) and subregion (bottom) maps. Boundaries are approximate, since they are based on BA footprints rather than hard geographical boundaries. Source: EPA & EIA / US Government work / Public domain.

utilities, in which case LMPs are not as easy to obtain. Fortunately, due to the increasing interconnectedness of U.S. electricity markets, LMPs for BAs in many of the unrestructured states are reported to and available from ISOs with interfaces to these BAs.[10]

We gather hourly, real-time market LMP data for 2015 for each U.S. state. The sources for the LMPs used in this study are shown in Table 4.4. We use load zone APNs where possible, although for most of the non-California BAs in WECC only generation APNs are available from CAISO. The selected generation-based and load-based APNs each represent an aggregation across fairly large geographic areas, and so, to some extent, they smooth out localized price spikes. However, there may be differences in these price types; in particular, an examination of a few zones for which both generation and load APNs are available in WECC shows that, while energy and congestion charges are generally the same, the loss component is—as we would expect—less for the generation APN than for the load APN. In typical circumstances, loss components are small (on the order of \$0-\$2/MWh) and (at least for the APNs in WECC) frequently negative. However, if inter-region LMP differentials are within the range of the loss component, then the use of generation-based APNs for some western states may cause those states to be favored against states using load-based APNs. For the ISOs, our average LMPs generally agree with the ISOs' reported averages for 2015 [223]. Additionally, the LMPs in Florida and Georgia are identical in the MISO data, even though they are designated with different nodes.

Figure 4.7 shows the general distribution of LMPs in each state, grouped by subregion. We observe that prices within the same subregion are correlated, which is expected given the operation of the electric grid. Some states generally have lower prices than others, and so we would expect a flat-rate, static approach to achieve savings vs. the baseline just by locating data centers in these lower-priced states. However, the temporal variation in LMPs means that such a static strategy is not likely to be dominant—that is, further savings can be achieved by leveraging hourly price differentials.

The boxplots are truncated at the high and low ends to make it easier to compare the "normal" LMP range of each state. However, there are many hours in which prices exhibit large positive or negative spikes, and Figure 4.8, which shows hourly LMPs for four different states in January and June, better exhibits this temporal volatility. We observe both geographic and temporal variability, with a winter heating effect in the Northeast in early January being particularly noticeable. Prices in New York and

[10] CAISO and MISO are particularly valuable for obtaining out-of-footprint LMPs, although MISO dropped 28 "second-tier" external pricing nodes (those that are not directly connected to the MISO system) as of 6/1/2016, so MISO will be a less useful source for LMPs in the Southeast going forward [222].

Table 4.4: Data sources for locational marginal prices.

ISO/RTO/BA	States	LMP Source
ISO-NE	CT, MA, ME, NH, RI, VT	ISO-NE [224]
NYISO	NY	NYISO [225]
PJM	DE, KY, MD, NJ, OH, PA, VA, WV	PJM [226]
MISO	AR, IA, IL, IN, LA, MI, MN, MO, MS, ND, SD, WI	MISO [227]
SPP	KS, NE, OK	Available from SPP [228], but obtained from MISO [227] due to easier interface.
ERCOT	TX	ERCOT [229]
CAISO/PGE	CA	CAISO [230]
WECC BAs: AZPS, PSCO, IPCO, NWMT, PNM, NEVP, PACW, PACE, BPAT, WACM	AZ, CO, ID, MT, NM, NV, OR, UT, WA, WY	CAISO [230]
SERC BAs: AEC, FPL, SOCO, CPLE, SC, TVA	AL, FL, GA, NC, SC, TN	MISO [227]

Massachusetts seem generally higher than in California and Texas in the winter, with the opposite being true in the summer—although no state is price-dominated in either month. Qureshi *et al.* [204] further demonstrate the variability of LMPs across space and time and its application to a data center load shifting strategy.

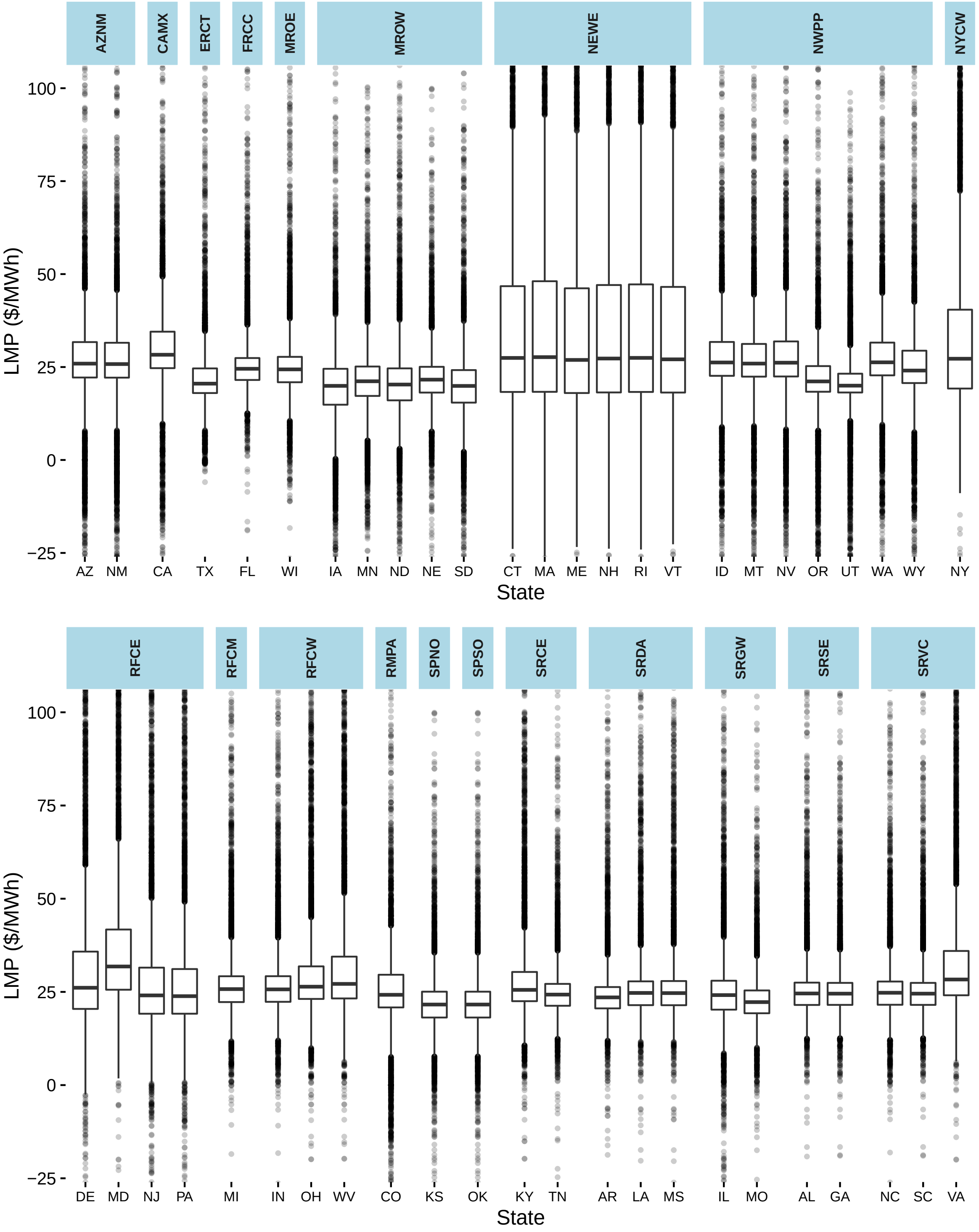


Figure 4.7: Boxplot of LMPs by state and subregion. Middle horizontal bar represents median; hinges are at 1st and 3rd quartiles; whiskers extend to 1.5 times the interquartile range past the upper and lower hinges. Y-axis is truncated at -$25 and $100/MWh for readability; there are many outliers beyond these limits.

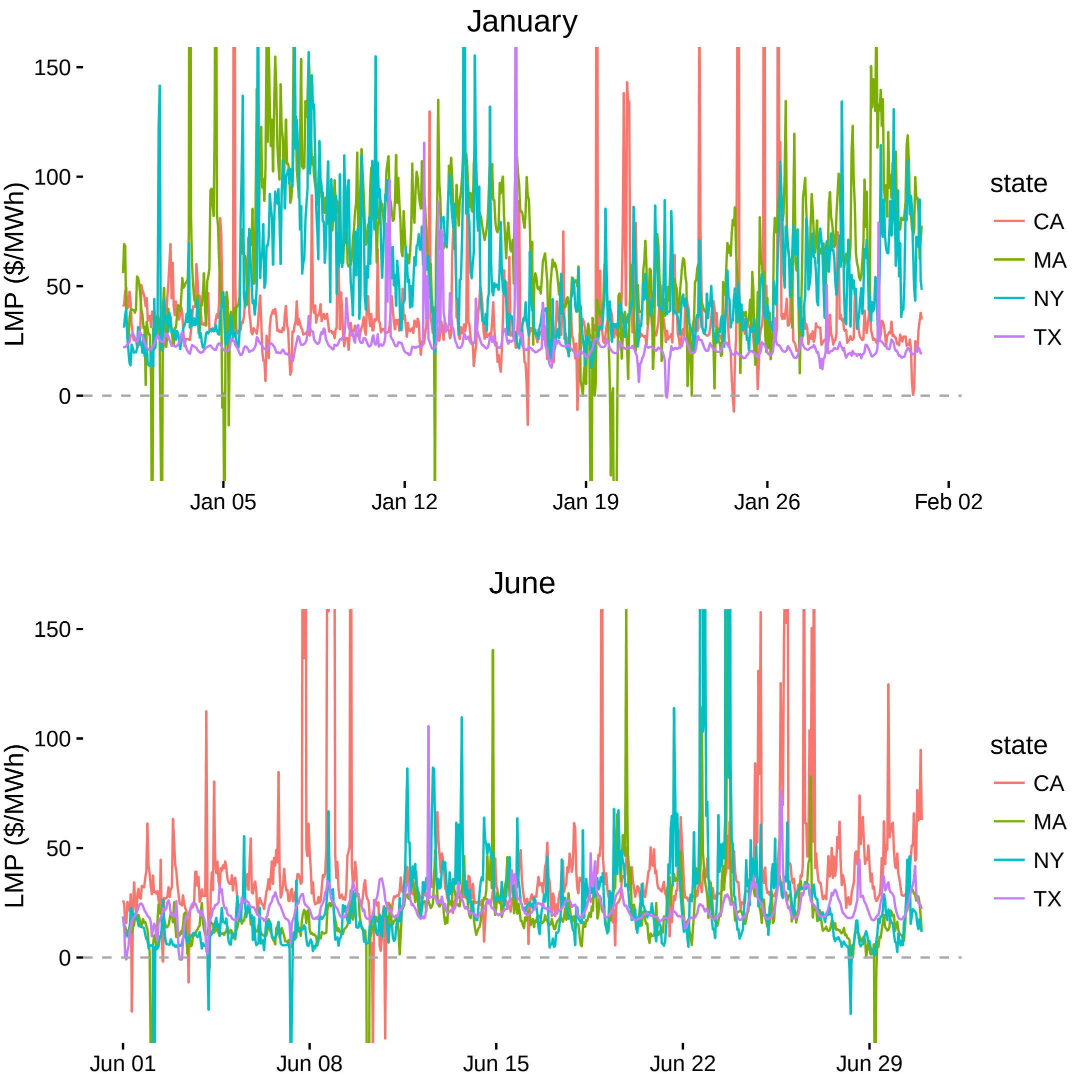


Figure 4.8: Hourly LMPs for selected states in January and June, 2015, show both geographic and temporal variability. Note: vertical axis truncated at [-$30, $150].

Beyond LMPs: wholesale and retail price adders

LMPs are at the heart of the load shifting strategy explored in this analysis. The temporal and geographical variability of real-time energy prices create the potential arbitrage opportunity. While looking only at LMPs may be sufficient to do a first-order comparison of different load shifting strategies, a more realistic analysis will convert the LMPs into the retail prices faced by the consumer by including various “adders” in the modeled price. There are two primary reasons to estimate these markups to the LMPs. First, these adders vary regionally. If they were constant nationwide, then we could perhaps ignore them; but they contribute to the differential pricing that an electricity consumer would face. Second, it will be useful for us to compare the load shifting strategies against a baseline strategy where the operator pays standard retail commercial or industrial electricity rates. Therefore, we need to convert electricity costs to a retail basis—a step that the earlier papers looking at data center electricity price arbitrage did not undertake.

While the most dynamic, energy (i.e., as set by the LMP)[11] is only one of several components in wholesale and retail prices. Because “electrons are not bound by contract, but instead obey the laws of physics,” [220] and because electricity is a public good, necessitating a stable, reliable grid, ISOs add charges for a variety of ancillary services (which maintain grid reliability), capacity planning, “uplift” or “make-whole price” (which reimburses generators for periods when they are forced to sell at a rate below their marginal cost of production), and other fees [231]. This fully-burdened wholesale price is sometimes called the “all-in” price. See Figure 4.9 for a more complete breakdown.

In PJM, the energy portion accounted for approximately 75% of wholesale costs from 2011-2014 [232]. The same was true in ISO-NE for 2015 [233]. In NYISO, 2014 energy costs were 77-83% of the total wholesale cost in all zones except for New York City, where rising capacity costs meant that energy accounted for only 62% of the wholesale cost [234]. These proportions are somewhat variable from year to year. In 2011, for instance, energy costs were 79% of the all-in price in NYC and 95% of the all-in price elsewhere [235]. In contrast to ISO-NE, NYISO, and PJM, these added costs are relatively small in ERCOT, CAISO, MISO, and SPP [236, 237]. We assume that wholesale adders in regions not governed by ISOs are also small.

Beyond adding these other wholesale charges to the LMP, we also need to account for retail adders: transmission and distribution costs as well as state and local taxes and other customer fees. However,

[11] Note the distinction between *energy* as a component of LMP, and *energy* within the scope of the wholesale or all-in price. In the latter case, *energy* usually refers to the entire LMP.

estimating retail electricity prices from wholesale prices is not straightforward. ICF International, under contract to the EPA, has a proprietary wholesale power market model that feeds a retail model. The documentation of the model notes several complicating factors, including the fact that retail prices are fundamentally different depending on whether the market is restructured: deregulated markets are estimated using a $\textit{wholesale} + \textit{transmission} + \textit{distribution}$ modeling approach, whereas regulated markets use a cost-of-service model [238]. We adopt the former approach, obtaining average transmission and distribution costs at the subregion level (See Figure 4.6, bottom) from the EIA [239, Table 55].

The EIA also reports monthly average retail prices at the state level separately for residential, commercial, industrial, and transportation sectors [240]. Typically, residential prices are the highest, followed by commercial, with industrial prices being the lowest.

Not included in our retail price estimates thus far, but included in average retail prices reported by EIA, are state and local taxes, demand charges, customer service charges, and other miscellaneous fees paid by end-use customers. These also vary by state and utility; New York, for example, has notoriously high electricity taxes that comprise as much as 25% of customers' bills, according to some estimates [241]. Given this broad heterogeneity, accurately modeling these charges using bottom-up data is a difficult task. Therefore, in order to compare our LMP-based prices to the EIA-reported retail prices on an equitable basis, we calculate the shortfall between our average retail price estimates and the EIA prices for each state.

Thus, our model of end-user electricity price is:

$$P = L + W + T + D + R \tag{4.12}$$

where P is the retail price, L is the LMP, W is the wholesale adder, T is the billed transmission charge, D is the billed distribution charge, and R is the retail adder. These components are further described in Figure 4.9. As outlined above, L comes from the online ISO data archives, W is an average value estimated from figures in the ISO's annual state-of-the-market reports, T&D are EIA-reported subregional averages, and R—lacking a primary data source—is the difference between EIA-reported state retail prices and our partial retail price estimate (i.e., $P = L + W + T + D$).

Figure 4.10 shows these price components by subregion and compares our estimates of retail rates (*sans* R) with the average commercial and industrial retail rates reported by EIA. In approximately half of the regions, our estimate matches the industrial retail price reasonably well. In the remaining regions,

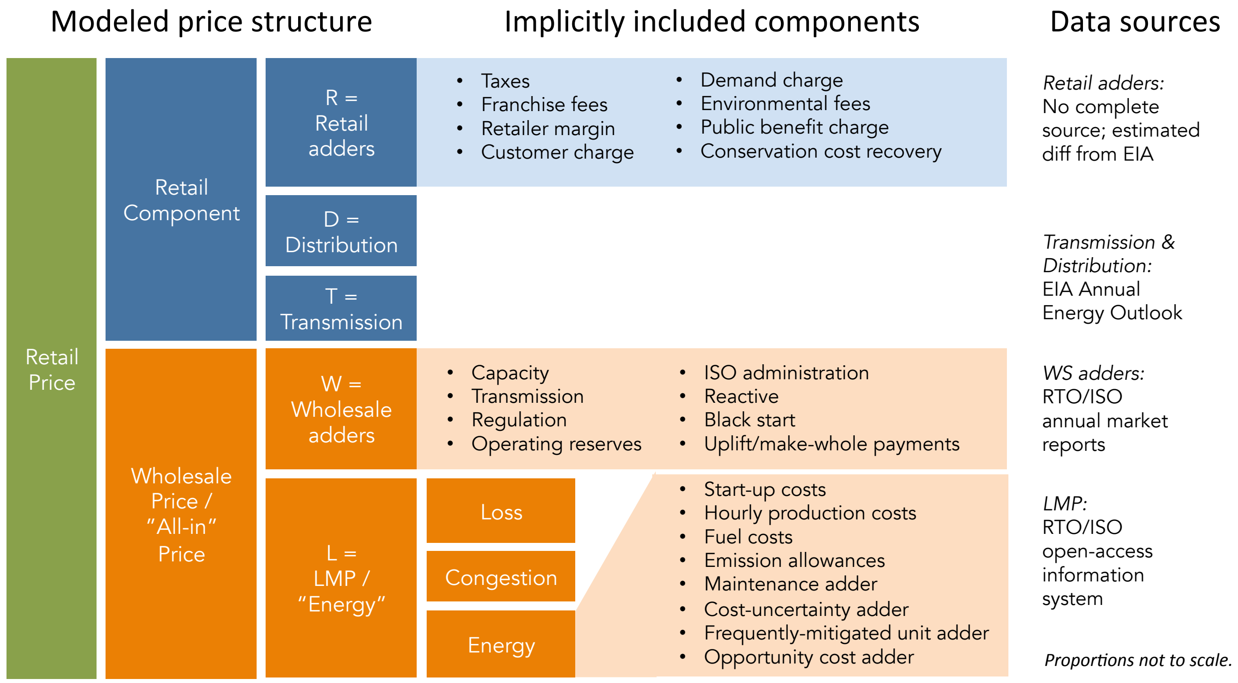


Figure 4.9: Components of retail electricity price. Not to scale. This is a notional breakdown; proportions vary from region to region, although the LMP is generally the largest component of the wholesale price, energy is the dominant component of the LMP, and fuel costs dominate the energy component of the LMP. The darkly shaded boxes ("modeled price structure") indicate price components explicitly included in our price model; the last column describes the sources for these data inputs. Lightly shaded boxes document charges that are implicitly included in the modeled components.

there is either a gap between our estimate and the industrial price, indicating the need to add R, or our estimate exceeds the industrial price. There are several possible explanations for the latter case: industrial electricity pricing is structured very differently than other rate schedules; industrial customers are large enough to negotiate better rates; it is cheaper for utilities to distribute power to a few large industrial customers than to many commercial and residential customers; the predictability of some industrial loads may mean that these customers can be billed based on the cost of base load generation; and state governments use low electricity prices as an incentive for attracting industrial jobs [242]. The reality of this gap is starkly illustrated in New York, where residential rates are among the highest in the country, but industrial rates are below the national average [243]. Because $T\&D$ values from EIA are averages across all sectors, it is certainly possible that our calculation overestimates these components for industrial users. Therefore, we allow the R term to be either positive or negative—that is, we can adjust our retail price estimate *up* to account for missing utility charges and *down* in regions where industrial rates enjoy a heavy discount.

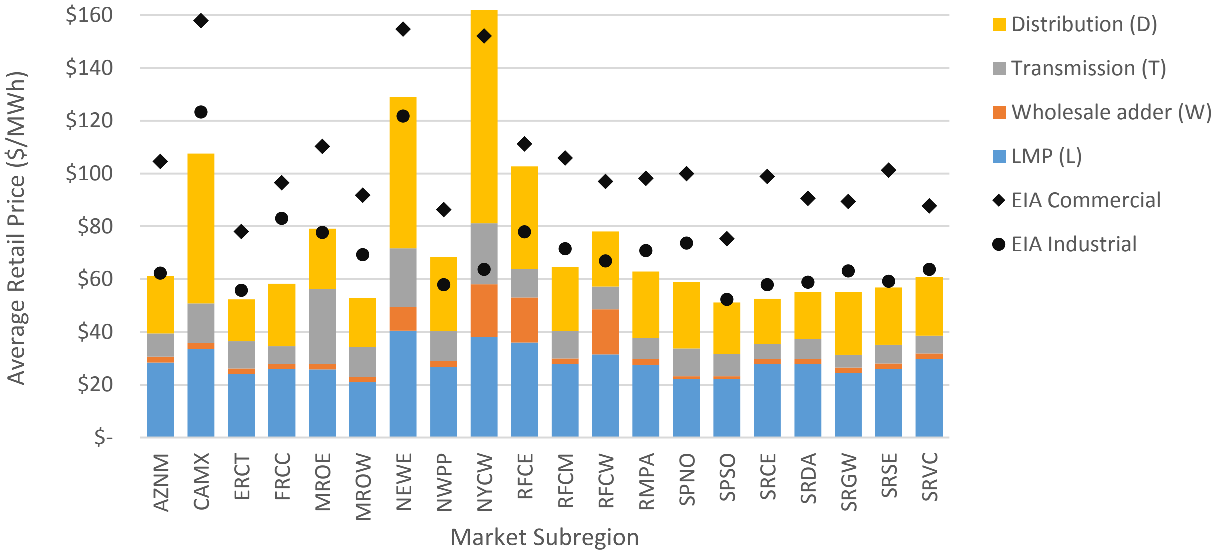


Figure 4.10: Regional retail commercial and industrial electricity price variation. The bars show our average estimated retail prices (without a retail adder) compared to the EIA-reported commercial and industrial prices for 2015. (See Equation 4.12.) EIA prices are load-weighted average prices aggregated from the state and sector tables of the Electric Power Monthly. *R* (not shown) is calculated as the average difference between the EIA prices and our estimates.

Customer exposure to dynamic prices

The load shifting strategy assumes that data center operators are exposed to the price volatility documented above. In practice, this tends to not be the case. CDNs like Akamai typically have their servers hosted in colocation centers, where they are billed a fixed charge for power capacity plus a flat metered rate (e.g., the commercial or—for hyperscale facilities—industrial rate offered by the utility) that accounts for the facility's PUE. That is, the energy consumed by Akamai's servers is scaled to account for the cooling and other energy overhead of the facility. Operators who own their facilities likely have bilateral power purchase agreements with utilities. There is such a wide variety of commercial and industrial electricity rate schedules that it is difficult to select a single representative billing model.

However, dynamic hourly pricing linked to wholesale markets, or RTP, is conceivable; indeed, a 2004 survey of RTP noted that over 70 utilities had such programs, with the first implemented in California in the 1980s [244]. Such rate structures are more likely offered to large industrial or institutional customers, though there are now also some residential programs [e.g., 245]. Many economists argue that linking retail and wholesale prices is an important step in making the electricity system more efficient, though there are a variety of mechanisms for establishing such a link [246, 247].

Note that LMPs can occasionally go negative. Negative LMPs occur in times of low demand when certain types of generators, such as nuclear base load, hydro, and renewables are willing to *pay* customers to consume their power for short periods of time, when the costs of curtailing generation have higher magnitude than this payment [248]. Consumers participating in hourly pricing programs indexed to the wholesale markets can expect to be paid for electricity consumption during these periods, though delivery and various other charges would still be assessed [245].

4.3.3 Damage data

Damages result from emissions of pollutants from power plants, including criteria air pollutants like nitrogen oxides (NO_x), sulfur dioxide (SO_2), and particulate matter (PM) as well as greenhouse gases like carbon dioxide (CO_2). As noted above, LMPs have a parallel construct in *marginal damages*. As different types of generators dispatch in response to changing demand, they not only change the marginal price of electricity, but they also change the marginal damages imposed on the public. However, whereas increasing demand should in theory always lead to an increase in marginal price, the direction of the marginal effect on damages could go either way, depending on the fuels of the plants involved. For example, if an intermediate coal plant is replaced by a natural gas peaker as the marginal generator, the marginal price will probably increase while the marginal damages will decrease. If, instead, natural gas replaces hydro as the marginal unit, then both price and damages will increase.

Estimation method

We obtain hourly power plant emissions data for 2015 from the EPA's Continuous Emissions Monitoring System (CEMS) [249], which contains hourly emitted volumes of NO_x, SO_2, and CO_2 from power plants as well as hourly generation load. We estimate $PM_{2.5}$ emissions by first dividing the annual $PM_{2.5}$ emissions from the National Emissions Inventory [250] by the annual generation of each plant from the CEMS data to find each plant's emissions rate and, second, multiplying hourly generation (again from CEMS) by this rate. For plants in CEMS but not in the NEI, we use the average $PM_{2.5}$ emissions rate for the plant's fuel type (coal, natural gas, or oil) as identified in eGRID [251], a listing of electricity generator characteristics.

We aggregate the data at the subregion level and thus have an hourly time series of fossil generation and emissions within each subregion. We can then calculate marginal emissions factors (MEFs) for each subregion, using a regression approach similar to that in Siler-Evans *et al.* [252], regressing emissions on generation separately for each season and hour of the day:

$$E = \beta_0 + \beta_1 G + \epsilon \qquad \forall\{p, e, s, h\} \tag{4.13}$$

where E is the emissions, and G is the fossil generation. β_1 is thus the marginal emissions factor (MEF), or the change in emissions associated with an incremental change in generation. We run this model separately over mutually exclusive subsets of the data partitioned along four dimensions, p, e, s, and h, where p is the set of pollutants, CO_2, SO_2, NO_x, and $PM_{2.5}$; e is the set of eGRID subregions;[12] s is the season, where winter is November – March, summer is May – September, and April and October are transition months; and $h = 0 \ldots 23$ is the hour of day. Thus, we obtain 5,760 separate MEFs—i.e., 24 *hours* × 3 *seasons* × 4 *pollutants* × 20 *subregions.*

Equation 4.13 estimates emissions factors. The next step is to translate these MEFs into dollar-valued marginal damages. For CO_2, we use a social cost of carbon value of $40/ton. For the criteria pollutants, the damage associated with a unit of emissions varies by location as a result of population density, environmental conditions, and other factors. Therefore, we translate emissions into damages *at the plant level*—i.e., prior to aggregating to the subregion level and running regressions. Thus, the regression is similar to Equation 4.13, but with damages instead of emissions as the response:

$$D = \beta_0 + \beta_1 G + \epsilon \qquad \forall\{p, e, s, h\} \tag{4.14}$$

where D represents the aggregate damages in a region. β_1 is now the marginal damage factor (MDF), or the change in damages associated with an incremental change in generation. We might think of the power plant emitting dollars (or, more appropriately, *negative* dollars) instead of pollutants and GHGs.

We use two different models to convert criteria air pollutant emissions, E, to damages, D. The AP2 model [253] is an integrated assessment economic model of U.S. air pollution that accounts for health effects, reduced timber and agricultural yields, reduced visibility, material degradation, and losses in recreation services that result from emissions from each source [254]. EASIUR is a reduced-form, regression-based estimation of outputs from higher-fidelity (but computationally expensive) chemical transport models [255]. In general, both models take emissions as inputs, use an air quality model to convert

[12] While New York City, Long Island, and Upstate New York are separate subregions, we treat the entire state as a single subregion, labeled NWYK

the emissions into geographic pollution concentrations, estimate the exposed populations in these geographic areas, calculate the effects on the exposed population using dose-response functions, and finally assess the value of those effects—e.g., using value-of-statistical-life figures to calculate damages associated with mortality.

The MDFs for each pollutant are summed within each hour to yield an overall MDF for fossil generation in each seasonal hour in each subregion:

$$\beta_{e,s,h} = \beta_{CO_2,e,s,h} + \beta_{SO_2,e,s,h} + \beta_{NO_x,e,s,h} + \beta_{PM_{2.5},e,s,h} \qquad \forall\{e,s,h\} \tag{4.15}$$

In addition to estimating the marginal damage factors, we also calculate the average damage factor (ADF) for each seasonal hour, subregion, and pollutant:

$$ADF = \frac{\sum D_{p,e,s,h}}{\sum G_{p,e,s,h}} \qquad \forall\{p,e,s,h\} \tag{4.16}$$

where D and G are damages and fossil generation, as above, and the 5,760 separate AEFs are calculated. The individual pollutant AEFs are summed to yield an overall AEF for each seasonal hour in each region, similar to Equation 4.15

Thus, we have four sets of damage factors: average and marginal, each estimated separately using EASIUR and AP2. See the appendix for more detail on this regression approach, other modeling approaches, and a comprehensive summary of the damage factors used in this analysis. We highlight some important aspects of these estimates below.

Implicit assumptions

The regression approach described above implies two key assumptions. First, our generation data are aggregated from fossil plants only and thus exclude nuclear and renewable generation sources, so these MEFs assume that the marginal generator is coal-, gas-, or oil-fired. Second, this approach does not account for imports and exports between regions. Therefore, we assume that demand in a particular subregion is met by generation in the same subregion. This assumption is the reason we calculate MEFs at the subregion rather than the state level; using larger geographic areas mitigates somewhat the import-export concern. Section A.3 in the appendix includes a short discussion of alternative approaches that might allow removal of these assumptions, but for this analysis they remain in place. As a result, our MDFs may be too high in regions and hours where incremental demand is met by non-fossil generation, either within the region or as an energy import.

Damage factor patterns

Damage factors vary from regionally, seasonally, and hourly. Figures 4.11 and 4.12 show the ADF and MDF estimates for the Eastern Mid-Atlantic states (RFCE) and California (CAMX), respectively. Damage factors in RFCE are universally higher than those in CAMX, while winter has a marked effect in the former but not the latter. California has minimal SO_2 compared to the East Coast. This is an extreme example: a load shifting solution with only these two regions would always favor California when external costs are the focus. Section A.5 in the appendix contains similar charts for all of the subregions (Figure A.17 – Figure A.35). In particular, see Figure A.36 for a side-by-side comparison of all subregions on a fixed scale.

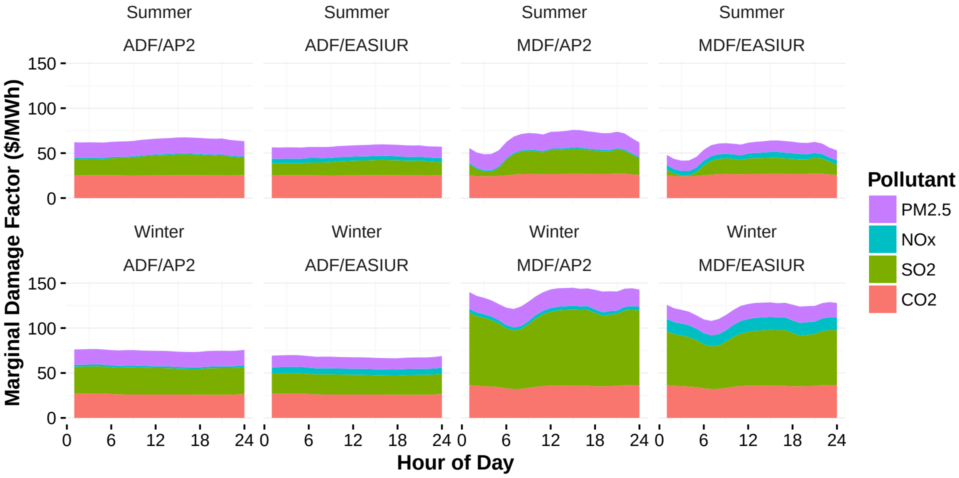


Figure 4.11: Seasonal hourly average and marginal damage factors for RFCE as estimated by AP2 and EASIUR.

Furthermore, the overall shape of MDF response to demand varies across regions. Subregions generally have the same daily demand profile, and thus similar generation profiles (Figure 4.13). (The fact that we are only tracking fossil generation, of course, means that our measure of electricity supply does not fully meet demand. Indeed, Figure 4.13 shows evidence of missing wind, solar, and possibly hydro in regions where those resources can often displace the need for fossil generation on the margin: AZNM, CAMX, and NWPP, for instance.) The different fuel mixes in each region mean that the daily MDF profile differs: for some regions (e.g., SRSO, Figure 4.14), the MDF generally rises and falls somewhat in sync with demand, while in others (e.g., RFCM, Figure 4.15), the MDF decreases during peak demand.

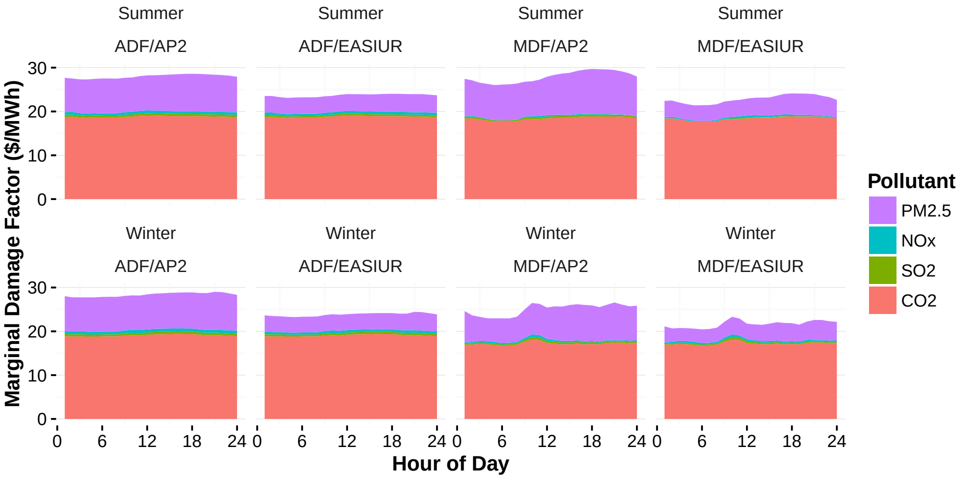


Figure 4.12: Seasonal hourly average and marginal damage factors for CAMX as estimated by AP2 and EASIUR.

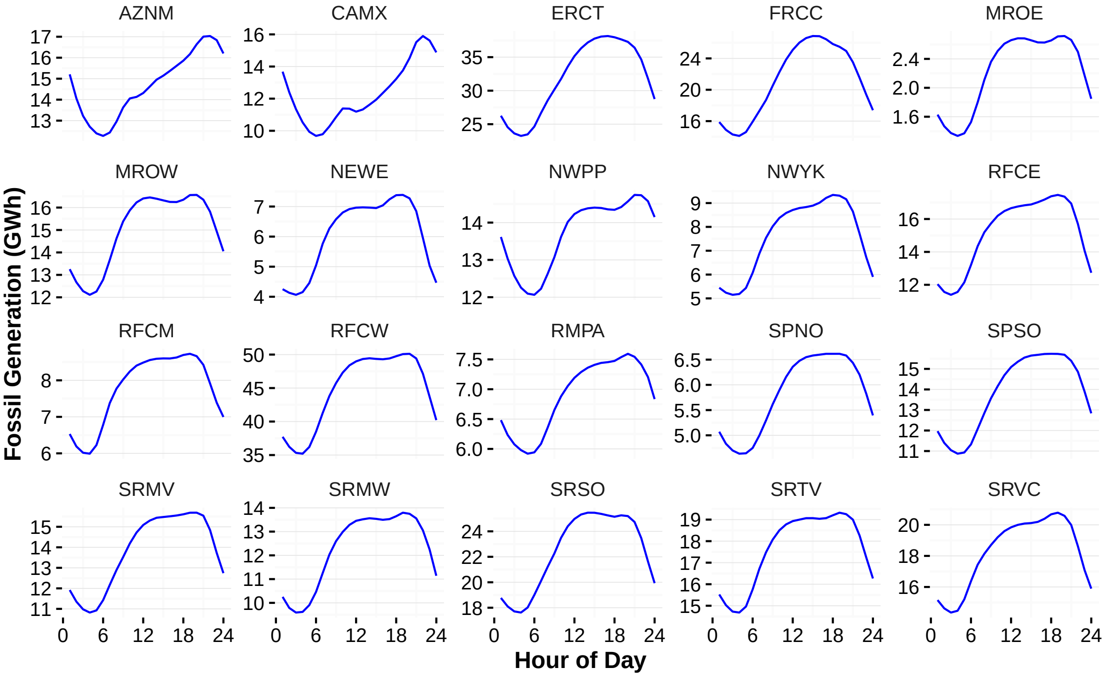


Figure 4.13: Hourly fossil generation by subregion. Note different y-axis scales. The hour of day is normalized to Eastern time, so the load profile is shifted later for regions in the Central (1 hour), Mountain (2 hours), and Pacific (3 hours) time zones. (See also Figure A.7 in the appendix.)

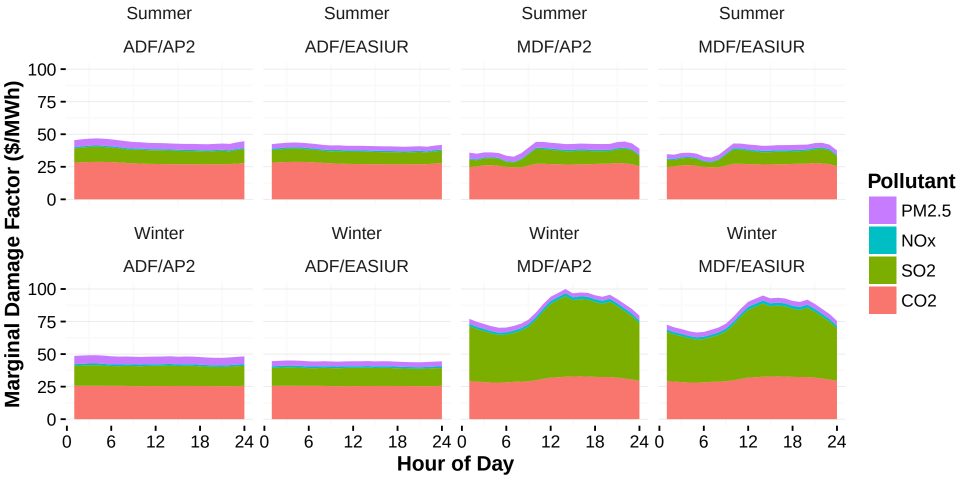


Figure 4.14: Hourly average and marginal damage factors, SRSO subregion.

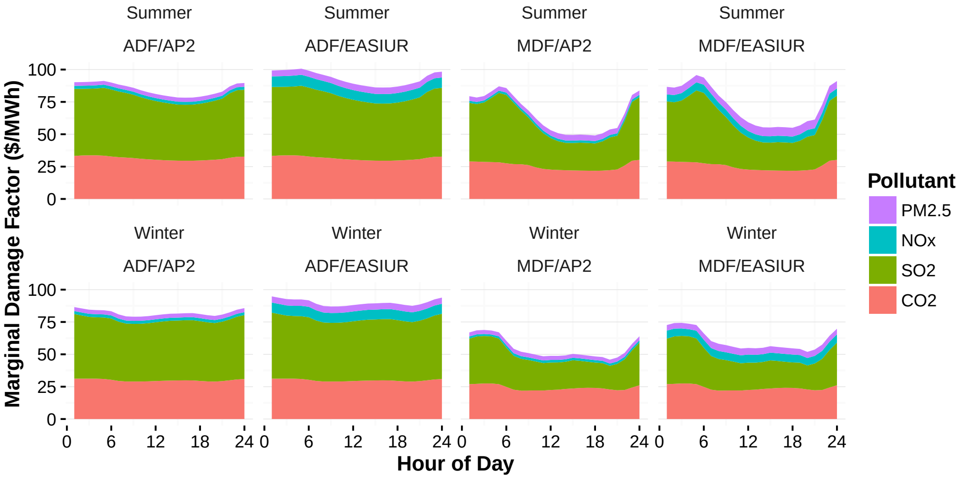


Figure 4.15: Hourly average and marginal damage factors, RFCM subregion.

4.3.4 Correlation between LMPs and damages

Before moving on to routing strategies, we briefly evaluate the relationship between the two cost measures. We know that, in normal market operations, LMPs increase with demand, and we also know that the marginal damage curve is more complex, as discussed above. Therefore, we expect some divergence between LMPs and MDFs: there should be some time periods where marginal private costs are increasing, but marginal external costs are decreasing. The degree of correlation (or anticorrelation) between these two factors will—in part—dictate the costliness of the tradeoffs involved in the different cost minimization strategies and whether or not a solution providing a “win-win” from both the public and private perspectives can be found. High correlation will make reducing external costs an easier sell to data center operators.

Computing the correlation matrix between LMPs and MDFs within each region shows no compelling evidence for correlation in these factors. (The raw correlation coefficients are reproduced in Tables B.1 – B.3 in Appendix B.) Most correlations are around zero, with some reaching as high as 0.2 in magnitude. Michigan, for instance, exhibits correlation of -0.22 between LMP and MDF. Some of the off-diagonal correlations are as high as 0.3 in magnitude (see Michigan’s MDF against LMPs in Wisconsin and Minnesota, for instance), but we cannot conclude definitively that these cross-boundary relationships result from interstate electricity transfer; they may simply be a result of similar generation profiles.

We are comparing *actual* LMPs to *estimated* MDFs. We can also compute correlation between LMPs and the “actual” damages[13] to see if our MDF estimation is hiding a relationship. In each hour, we compute *damages*/*generation* and calculate the correlation with LMPs. The results still show very little correlation between these two cost measures, with the exception of the Northeast, where the correlation coefficient is as high as 0.7.

However, even if hour-to-hour damages and prices are not correlated within a region, there may be ordering *between* regions; that is, some regions may generally have higher damages and prices than others, or might consistently have high prices and low damages (or *vice versa*). State average LMPs and average MDFs have a relatively small correlation coefficient of about 0.2, and Figure 4.16 shows that the sets of states with favorable damages and favorable prices are somewhat different. Based on these results, we expect minimizing private costs to yield a different routing strategy than minimizing external costs.

4.3.5 Data summary

The key data values and assumptions described above and used as inputs to the model (described in the next section) are summarized in Table 4.5. There are three levels of electricity prices: commercial, industrial, and LMP. These prices can be either flat, varying only regionally, or dynamic, varying both regionally and hourly. Flat commercial and industrial prices come from EIA; the flat LMP rate is the state

[13] Really, actual *emissions*; the translation to damages using AP2 or EASIUR is still an estimate.

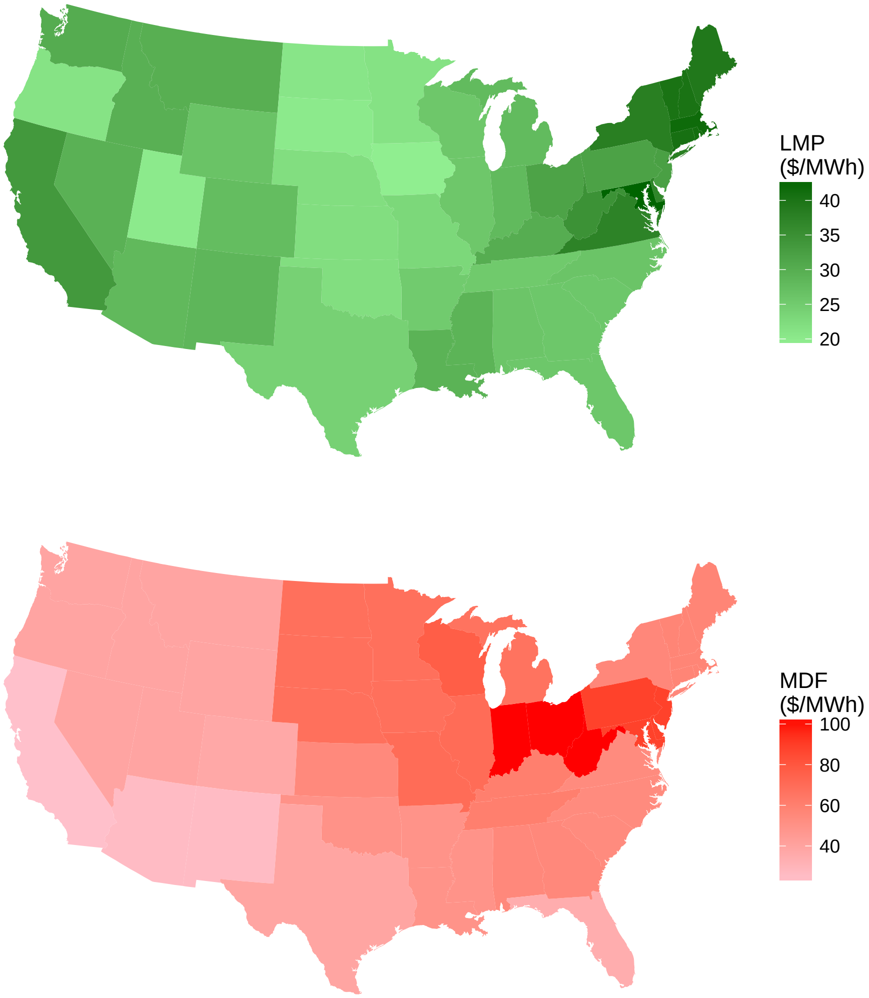


Figure 4.16: Heat maps of state average LMPs (top) and MDFs (bottom), which show differences in a state's cost favorability depending on whether the private or external cost metric is used.

average LMP for 2015. LMPs are the basis for the dynamic rates; dynamic commercial and industrial rates are estimated by supplementing LMPs with various price adders, as described in Section 4.3.2. We report results for the LMPs alongside the retail prices in most cases; however, we note that the LMP pricing scenario is unlikely to be a realistic representation of electricity costs. The CDN operator will most likely face a retail rate, regardless of whether it is dynamic or flat.

We use two types of damage factors, average and marginal, as estimated by two different models, AP2 and EASIUR. In theory, the marginal factors are the more appropriate metric for this analysis, since we are interested in incremental effects.

Table 4.5: Summary of input data.

Input	Basis of Estimate	Source	Variation
Traffic (Gb/hr)	Akamai U.S. web traffic	Node-client loads from Akamai, scaled to total U.S. load	Hourly and regional
Electricity price ($/MWh)	LMP	LMP directly from wholesale markets	Hourly and regional
	LMP-based commercial	Retail price as estimated by supplementing LMP with commercial adders	Hourly and regional
	LMP-based industrial	Retail price as estimated by supplementing LMP with industrial adders	Hourly and regional
	LMP flat rate	Average LMP	Regional
	Commercial retail	Flat rate from EIA	Regional
	Industrial retail	Flat rate from EIA	Regional
Damages ($/MWh)	Marginal/EASIUR	Marginal damages estimated by EASIUR model from CEMS emissions and fossil generation	Hourly and regional
	Marginal/AP2	Marginal damages estimated by AP2 model from CEMS emissions and fossil generation	Hourly and regional
	Average/EASIUR	Average damages estimated by EASIUR model from CEMS emissions and fossil generation	Hourly and regional
	Average/AP2	Average damages estimated by AP2 model from CEMS emissions and fossil generation	Hourly and regional
Energy factor (kWh/GB)	0.01 kWh/GB	Parameter	None (Constant)

4.4 Model results: electricity cost savings estimates from load shifting

4.4.1 Establishing a baseline

The traffic loading for the baseline strategy using Akamai's actual routing is shown in Figure 4.17.

We also outlined a potential alternative baseline strategy based on proximal routing. We saw above that Akamai's routing does not always favor the closest data center, and, here, we briefly investigate how closely a purely proximal optimization matches Akamai's actual routing.

Under proximal routing, electricity costs and damages are within 4% of those in the baseline. Figure 4.18 shows the difference in traffic loads, in percentage-*points* between the proximal strategy and Akamai's strategy. The strategies largely match, with the exception of a handful of Akamai's main hubs: the proximal strategy underloads California, Illinois, and Virginia each by about 5% of the traffic, while

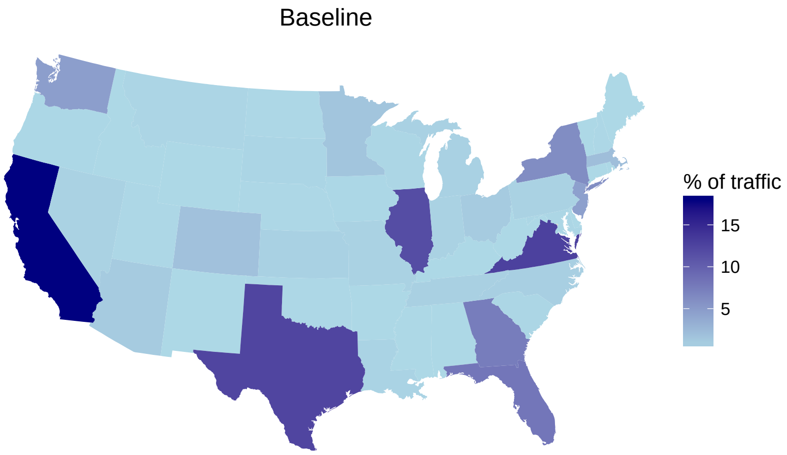


Figure 4.17: Heat map of traffic load under Akamai baseline routing.

overloading Washington and Florida by 2-3% of the traffic. The differences suggest that there are prioritized regional "hubs" in the CDN network, and the proximal strategy would not account for this priority.

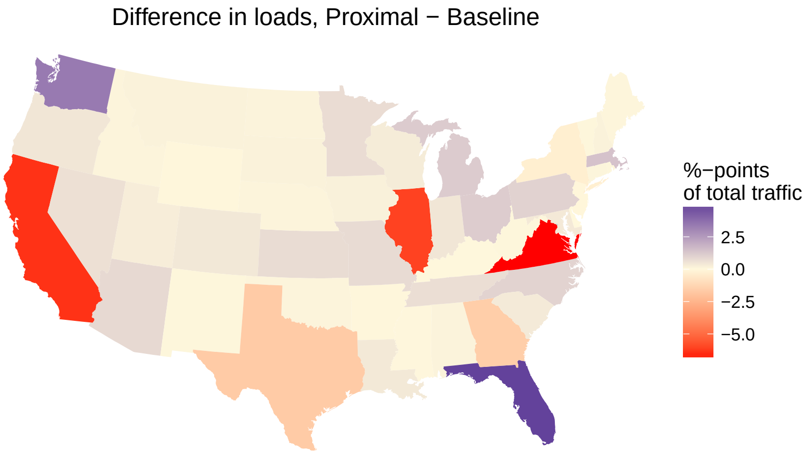


Figure 4.18: Heat map of difference between proximal and actual baseline routing. Differences are under- or over-allocation of traffic by the proximal scenario, measured in percentage-points of total traffic.

We use the simulated actual routing baseline for comparisons in the remainder of this analysis. However, the proximal routing strategy could be used as an approximation of a CDN routing algorithm in the event that real traffic routing data are not available—particularly if the focus is on electricity cost estimation rather than accurately representing individual node loads.

4.4.2 Minimizing private costs

Under the cost minimization strategy, the total price paid for electricity is minimized and external damages are tracked but do not have any weight in the objective function. As discussed above, we use six different pricing scenarios and assess the impact of each on both private costs and external damages, comparing the absolute differences in a series of bar charts, percentage savings in Table 4.6, and traffic allocation to nodes in a series of heat maps. When looking at the heat maps, it is important to keep two things in mind. First, they are snapshots in the dynamic pricing cases, and will change hour-to-hour in these scenarios. Second, node loads are jointly determined by price and node capacity, so a node with large capacity (e.g., California) may handle a lot of traffic even if it does not have the lowest electricity price in the circumstance where cheaper nodes are relatively small compared to the total traffic on the network. Thus, the "hottest" node on the map may not have the absolute lowest price.

Effect on private cost

First, we examine a flat-rate scenario using EIA prices in which prices are temporally static but regionally variable—that is, the network operator can allocate load to the region with the cheapest constant rate. In practical terms, this optimization could be run once per month or year, during which time the ranking would not change. This strategy should yield electricity savings compared to the baseline by making heavier use of data centers in regions where electricity is cheaper. Next, we repeat the optimization using dynamic hourly pricing, and we expect to achieve a larger percent savings, since now temporal variability provides an additional degree of flexibility in selecting data centers.

Figure 4.19 compares the private cost savings achieved under these different price scenarios. Serving load from the cheapest regions yields substantial savings in the flat-rate case, with dynamic price optimization yielding a smaller amount of further savings. The difference between the baseline and flat-rate costs in the LMP case is smaller than in the retail cases because minimizing only LMPs does not take advantage of regional differences in the various price adders that constitute the final retail rate.

Figures 4.20 and 4.21 show the traffic distribution among the nodes under LMP, commercial, and industrial flat-rate pricing and dynamic pricing, respectively. We note that in the retail rate cases, the loadings look similar between the flat-rate and dynamic pricing scenarios; this similarity is not surprising, since states with generally low average prices must by definition tend to have lower marginal prices,

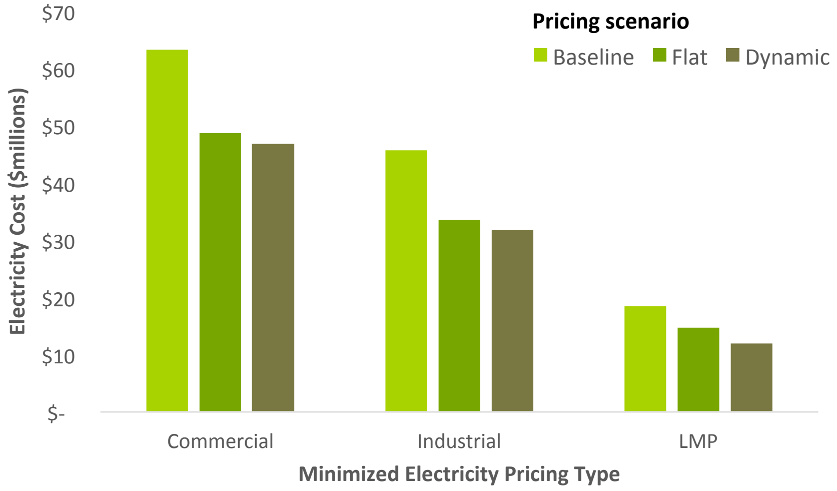


Figure 4.19: Comparison of minimized private costs under different pricing structures.

and since the EIA flat-rate prices were used to estimate the "adder" for the dynamic (LMP-based) commercial and industrial prices.[14] As discussed in reference to Figure 4.19, the dynamic optimization is simply "fine-tuning" the generally good flat-rate routing to extract further savings. Texas and Illinois are generally favored, while Georgia and New York are prioritized under industrial pricing and Virginia is prioritized under commercial pricing. Texas consistently handles on the order of 30% of all the traffic in these scenarios, indicating that tends to have the lowest price among the large hubs. The LMP model shows a greater degree of difference compared to the flat-rate pricing, making use of California in particular.

[14] See Section 4.3.2 for details.

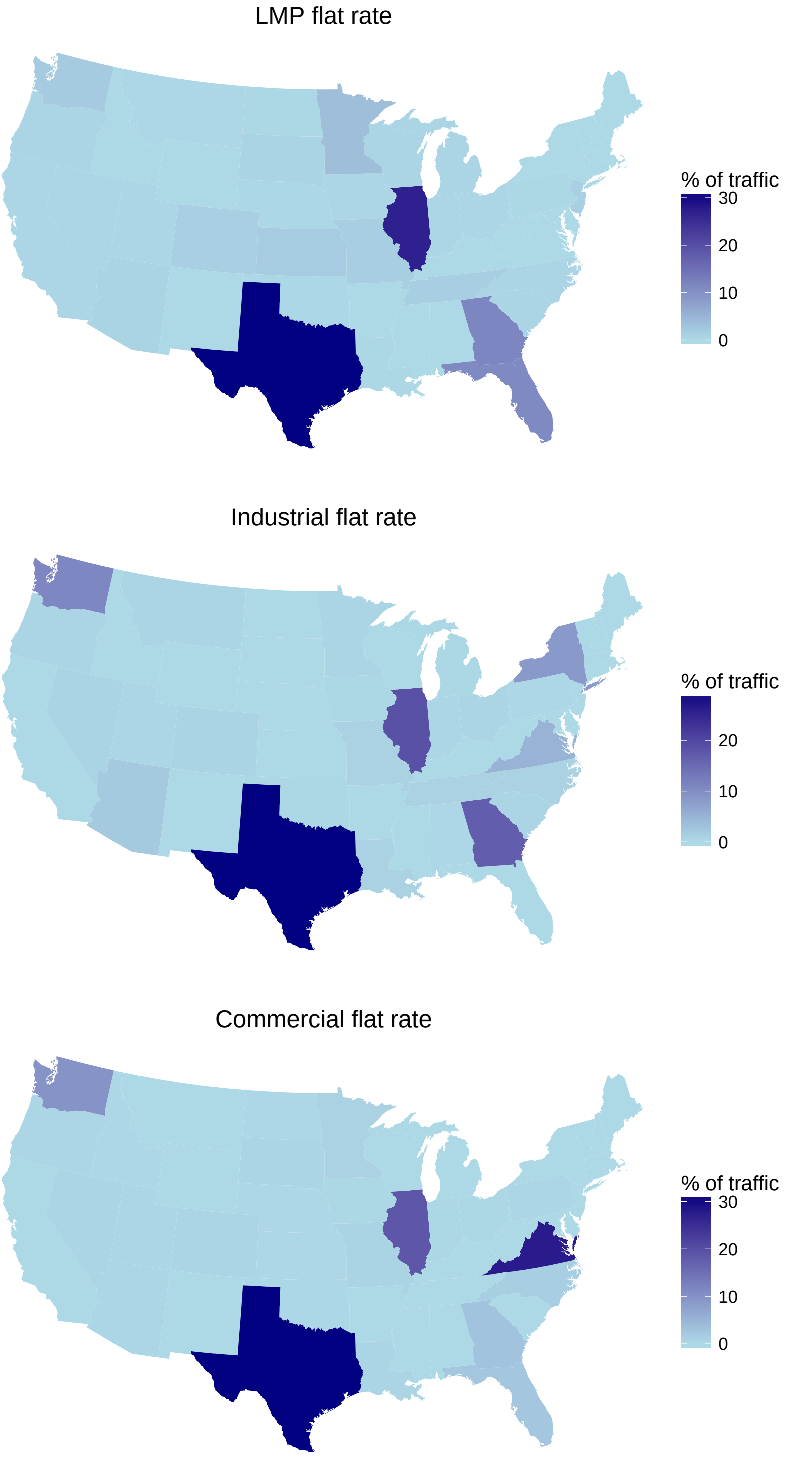
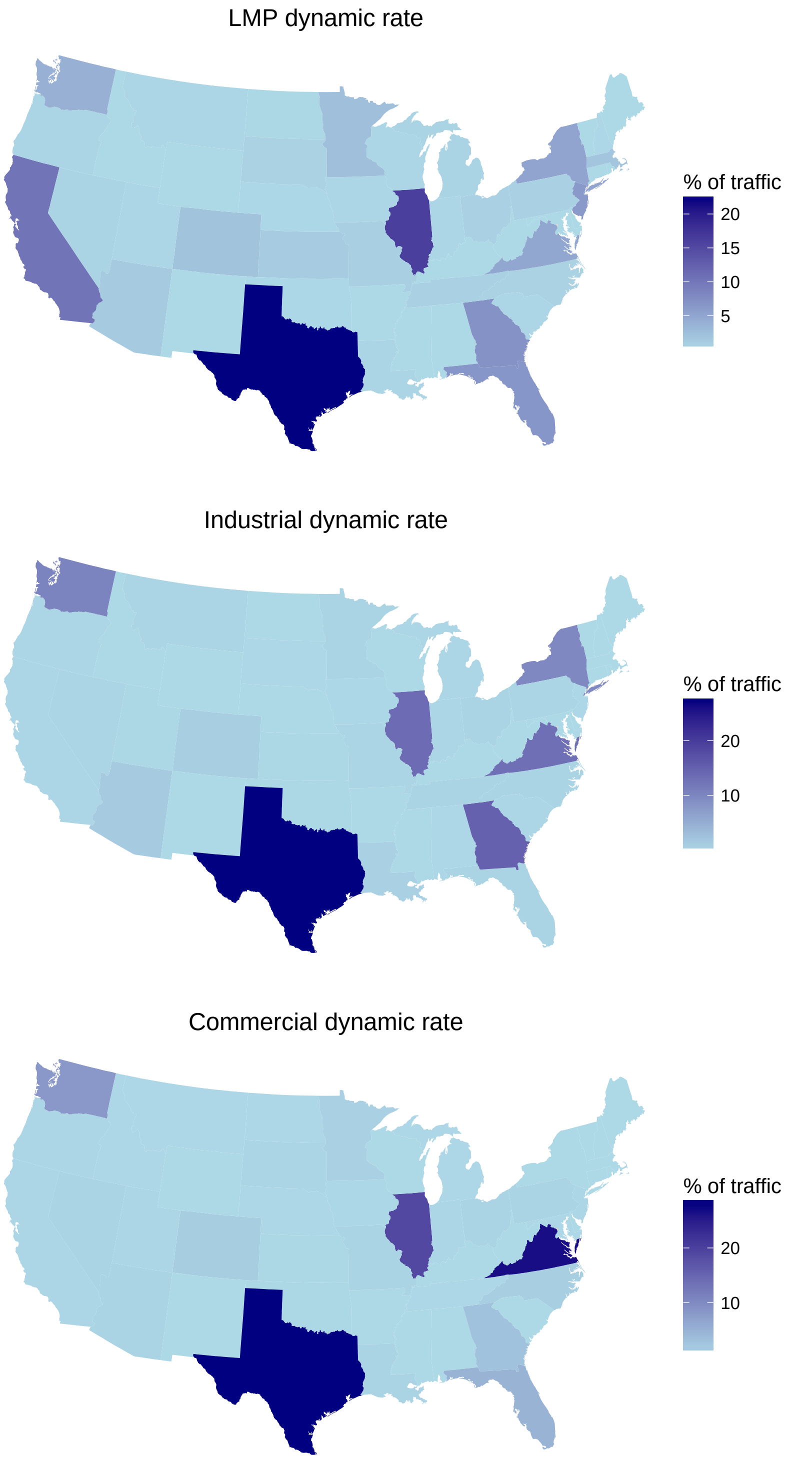


Figure 4.20: Heat maps of traffic load under under private cost minimization with average LMP (top), industrial (middle), and commercial (bottom) flat-rate pricing. Prices vary regionally, but not hourly, so in general these maps will be largely the same throughout the year. (There is some price variation month-to-month.)


LMP dynamic rate

Industrial dynamic rate

Commercial dynamic rate

Figure 4.21: Heat maps of traffic load under private cost minimization with LMP (top), industrial (middle), and commercial (bottom) dynamic pricing. Prices vary regionally and hourly, so these maps will look different hour-to-hour.

Effect on external cost

We now turn to the effect of the private cost minimization strategy on external damages. Figure 4.22 shows the external damages associated with minimizing each of the six different price scenarios as estimated by the four different damage models, with the baseline damages shown as a threshold. We can see that the damage models give qualitatively similar results: minimizing private costs increases external damages by between 2 and 10% in the retail price scenarios. Under LMP pricing, external damages increase by as much as 20%.

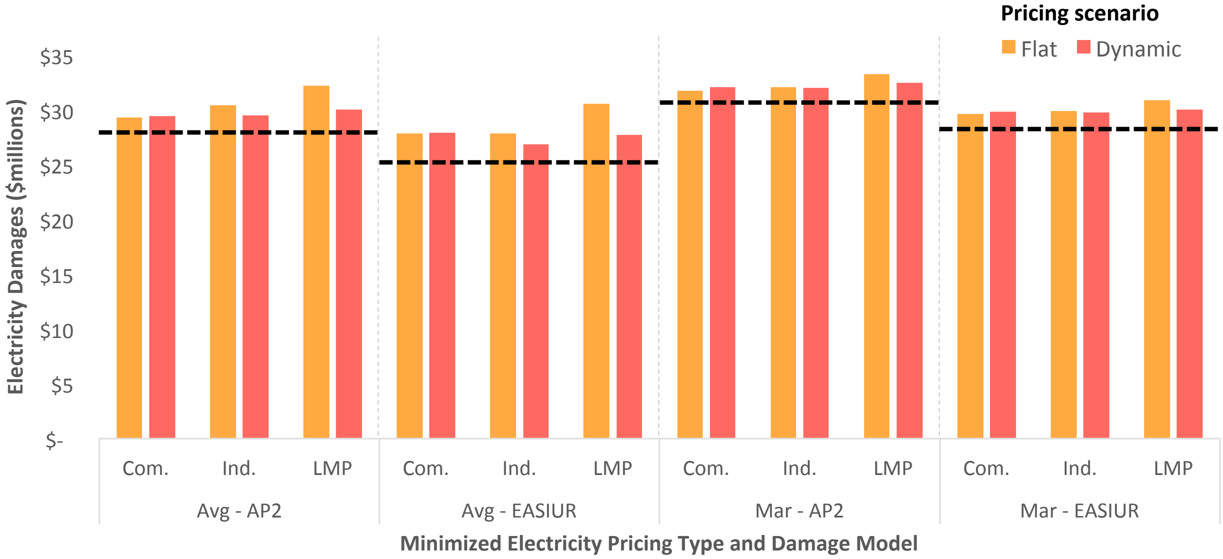


Figure 4.22: Effect of private cost minimization on external damages as estimated by four damage models under six electricity price structures. Horizontal dashed lines in each group represent the damages as estimated in the baseline routing case.

Savings summary for private cost minimization

Table 4.6 summarizes the relative private cost savings and damage impacts associated with minimizing each of the pricing scenarios. Under the retail prices that more realistically reflect the rates data centers are likely to face, we expect private cost minimization to reduce the power bill by a quarter if flat-rate pricing is used; when dynamic prices are added, then the savings increase to a third. Focusing on power bill minimization leads to an increase in external damages of 2% to 9%, depending on which damage model is used.

Table 4.6: Estimated savings and impacts under the private cost minimization strategy. The minimized cost is in blue; the right four columns show the side effects on external damages, estimated as both average and marginal factors using the AP2 and EASIUR models.

Price model	**Private savings**	**External savings (Average)**		**External savings (Marginal)**	
		AP2	*EASIUR*	*AP2*	*EASIUR*
Flat (LMP)	16%	-15%	-20%	-7%	-8%
Flat (Industrial)	27%	-9%	-9%	-3%	-5%
Flat (Commercial)	23%	-5%	-9%	-2%	-4%
Dynamic (LMP)	35%	-7%	-8%	-5%	-5%
Dynamic (Industrial)	32%	-6%	-5%	-3%	-4%
Dynamic (Commercial)	27%	-5%	-9%	-3%	-4%

4.4.3 Minimizing externalities

Figure 4.22 showed that minimizing costs associated with retail prices yielded a small to moderate increase in external damages. We expect that a strategy that minimizes external costs will show a potential for large additional reductions in external damages while generally increasing the power bill for the CDN operator.

Effect on external damages

As previously discussed, damages are estimated using four valuation methods. Figure 4.23 shows the substantial damage avoided by the damage minimization strategy, which ranges across the methods from 30% to 40%. Note that the baseline damages in each case represent the same load allocation; the differences reflect variation between marginal and average emissions factors and AP2 and EASIUR model outputs. Comparing the baseline figures, we can conclude that EASIUR generally provides slightly lower damage estimates than does AP2 and that average damages are generally lower than marginal damages for the same generation profile. The optimizer is able to erase the average-marginal difference through load shifting (the average and marginal damages are very close within each model), but the difference between AP2 and EASIUR remains.

Figure 4.24 shows a map of the traffic load allocation using minimized marginal damages for each damage model. California is heavily favored, handling almost half of the traffic in all cases, due to low damage factors resulting from a high proportion of natural gas generation as well as renewables penetration. EASIUR then prioritizes Texas and Florida, while AP2 prefers Virginia over Texas in the average case. We again note that traffic volumes are a function of both the minimized cost and capacity, so large nodes

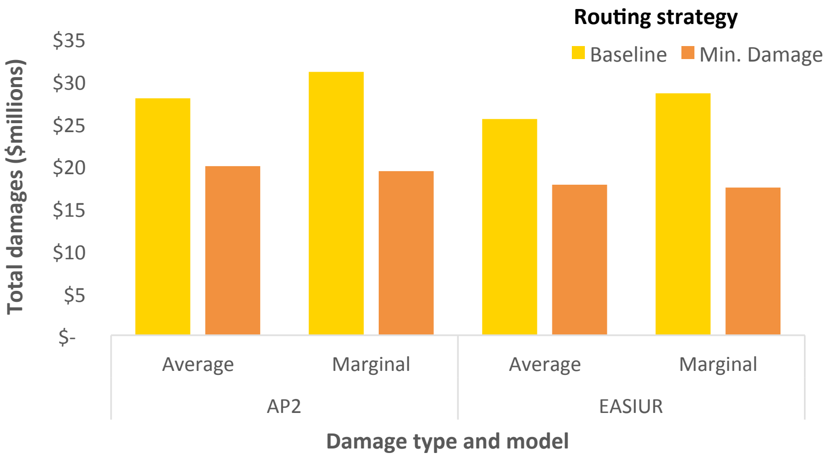


Figure 4.23: Comparison of minimized externalities under different damage models. Note that the baseline bars all represent the same traffic load allocation, while the min. damage strategy may reflect different load allocations.

tend to capture most of the traffic even if smaller states have lower cost, as those smaller states fill up. Nonetheless, clear differences between the damage minimizing and private cost minimizing strategies can be readily seen by comparing these maps with those in Figure 4.21. Illinois and Virginia, for instance, are heavily used, while California is avoided under private cost minimization, whereas California is heavily used and Illinois and Virginia are avoided under damage minimization.

Effect on private cost

This damage avoidance comes at a cost. Figure 4.25 shows the increased private electricity costs associated with minimizing damages by plotting private costs against the baseline threshold (dashed black line). The dotted blue line shows the minimized private costs (Figure 4.19) for reference. The increase vs. the baseline is relatively consistent across the damage models and for dynamic and flat-rate pricing: around 20% under industrial retail pricing, 15% under commercial retail pricing, and 0–5% under pure LMP pricing.

Savings summary for external cost minimization

Table 4.7 summarizes the avoided damages and increased private costs associated with this strategy. Under external cost minimization, damages can be reduced by 30%–40%, while the anticipated increase in the power bill is 15% under commercial pricing and 20% under industrial pricing.

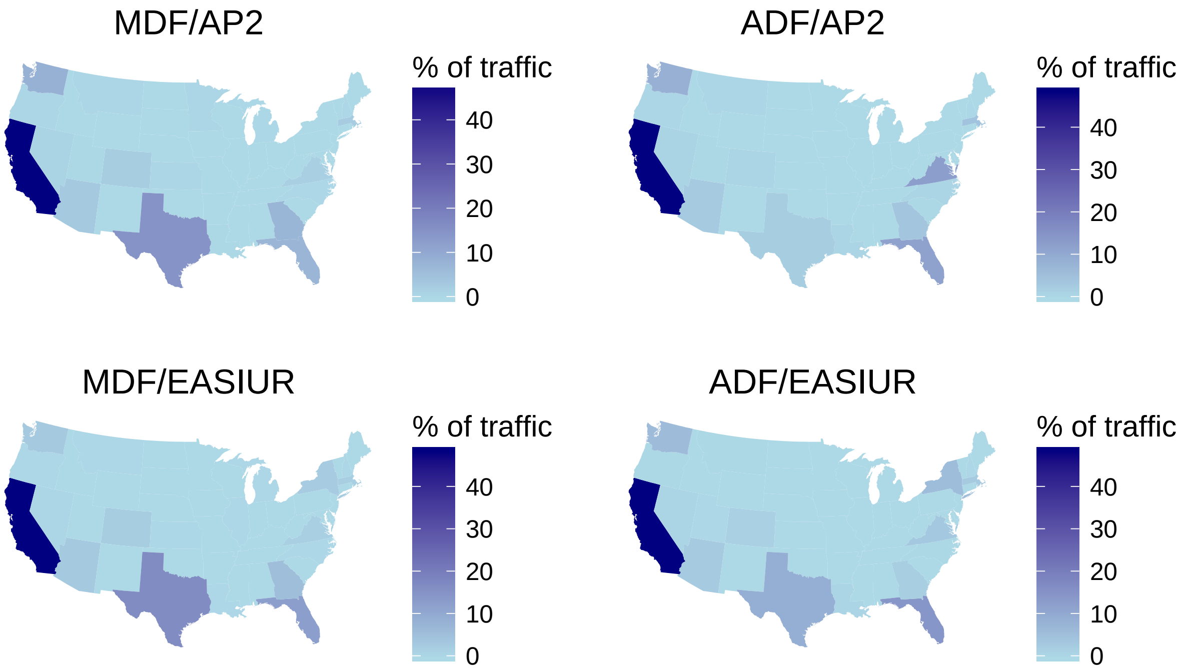


Figure 4.24: Heat map of traffic load under damage minimization under the four different damage models.

While we are focused on cost minimization, we can also evaluate the effect of load shifting on avoided pollutants. Table 4.8 shows the mass of pollutants associated with electricity generation under each scenario using marginal emissions factors at each node as calculated in Equation 4.13.

Table 4.7: Estimated savings and impacts under the damage minimization strategy. The minimized cost is in blue; the rightmost columns show the side effects on private costs, estimated under six different price scenarios. *Com* = commercial retail pricing; *Ind* = industrial retail pricing.

Damage model	**External savings**	**Private savings (Flat)**			**Private savings (Dynamic)**		
		LMP	*Ind*	*Com*	*LMP*	*Ind*	*Com*
Average (AP2)	29%	-7%	-23%	-16%	-4%	-21%	-15%
Average (EASIUR)	31%	-4%	-22%	-18%	-3%	-21%	-18%
Marginal (AP2)	37%	-1%	-18%	-14%	1%	-17%	-13%
Marginal (EASIUR)	39%	-1%	-20%	-16%	0%	-19%	-15%

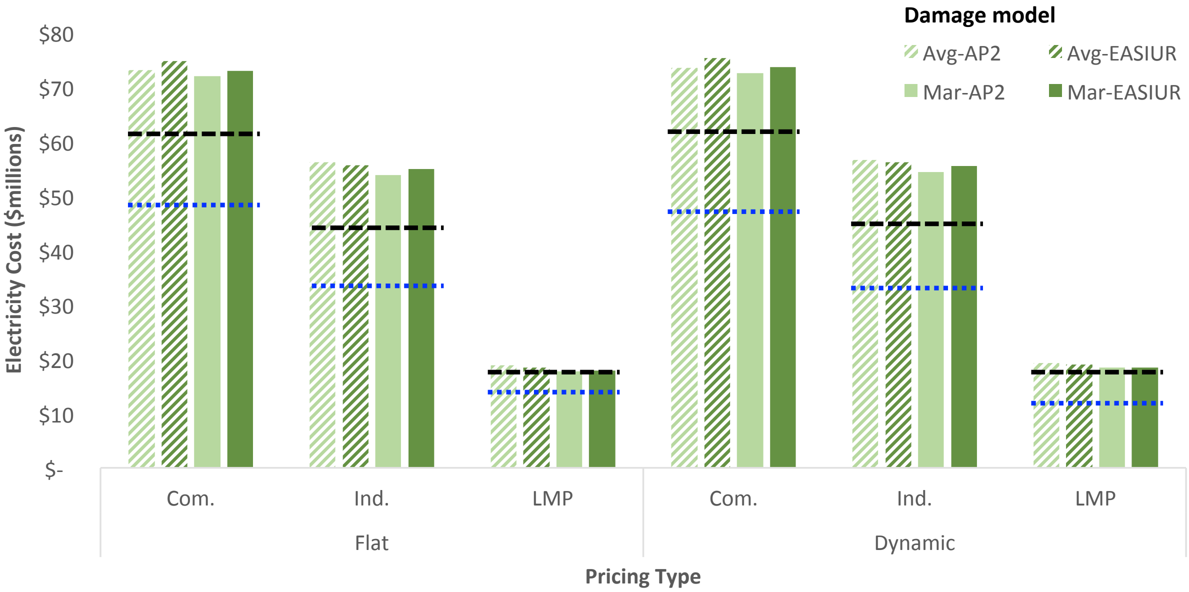


Figure 4.25: Private costs under the damage minimization strategy as estimated by six pricing models. Black horizontal dashed lines show the cost for the baseline routing strategy according to each pricing type; blue horizontal dotted lines show the minimized private cost. (See Figure 4.19.)

Table 4.8: Emitted pollutants associated with electricity generation under different load shifting strategies.

Scenario	**CO_2**		**SO_2**		**NO_x**		**$PM_{2.5}$**	
	(K metric tons)		*(metric tons)*		*(metric tons)*		*(metric tons)*	
	Emitted	vs. Base	Emitted	vs. Base	Emitted	vs. Base	Emitted	vs. Base
Baseline	375		281		204		41	
Min private	398	+23	347	+66	215	+11	38	-3
Min external	323	-52	144	-137	158	-46	35	-6

4.4.4 Pareto-optimal tradeoffs: results from joint optimization

We now have an initial sense of the tradeoff between private and external costs. The previous results suggest that minimizing either private costs or damages can reduce the targeted cost by about a third. The side-effects of the different strategies do not appear to be as balanced, however: minimizing private costs in the retail pricing scenarios causes a simultaneous increase of no more than 10%, whereas minimizing damages increases the CDN's electricity bill by 15–20%.

As noted above in Section 4.3.5, the LMP scenarios (the results of which fall outside the ranges just quoted in some cases) are probably not realistic. Having explored the impact of different pricing scenarios and damage models above, we adopt *dynamic industrial retail pricing* and *EASIUR marginal damages* as the “main” inputs used to explore model variations moving forward.

To further investigate the nature of the tradeoff between private and public cost minimization, we can plot the Pareto frontier by using different values of w from 0 to 1 in the objective function, Equation 4.5. When $w = 0.5$, private and external costs are weighted equally, and total costs are minimized. Different weightings of w move away from total cost minimization (assuming the frontier is nonlinear) but reflect other valid ways decisionmakers might prioritize costs. The shape of the Pareto frontier indicates how closely the joint optimization approaches the benefit of either individual minimization strategy. If there is a sharp “knee” in the curve (i.e., the frontier is very convex), then a “win-win” solution exists that will achieve a large proportion of both the potential private and external benefit. Alternatively, if the Pareto frontier is linear, then the tradeoff between external and private costs is greater, and it is impossible to capture a majority of the available savings in both metrics simultaneously.

Figure 4.26 shows the Pareto frontier for the joint minimization of private and external cost savings. Total costs are minimized where the isocost dashed line is tangent to the curve. At this point ($w = 0.5$), 85% (27 of 32 possible percentage points) of the total possible private savings and 34% (13 of the possible 39 percentage points) of the total possible external savings are achieved. These results confirm the earlier observation that external savings come at a somewhat disproportionate tradeoff to private savings—achieving $1 of external savings, on average, sacrifices more than $1 in private savings.

At the same time, the shape of the curve indicates a range of solutions that might be acceptable to both a private- and external-cost minimizer. Within the range $0.2 \le w \le 0.8$, savings relative to the baseline are simultaneously available to both public and private stakeholders. A map of the traffic loadings at $w = 0.5$ is shown in Figure 4.27.

Ultimately, the static maps presented here do not really provide the entire picture, which is dynamic in time as well as space. Figure 4.28 shows three time series of node loadings during the month of January, corresponding to the private, external, and total cost minimization strategies. These plots show more clearly than the maps the proportion of traffic allocated to each node and any variation over time. We observe the damage minimization strategy makes heavy use of California (the orange volume near the base of the plot), while the private cost minimization strategy makes heavier use of Texas and Virginia (the purple areas near the top of the plot). The total cost minimization strategy is a mix of the two, though in keeping with the finding that the tradeoff between external and private savings is not balanced, the strategy has more in common with private cost minimization than damage minimization. It makes

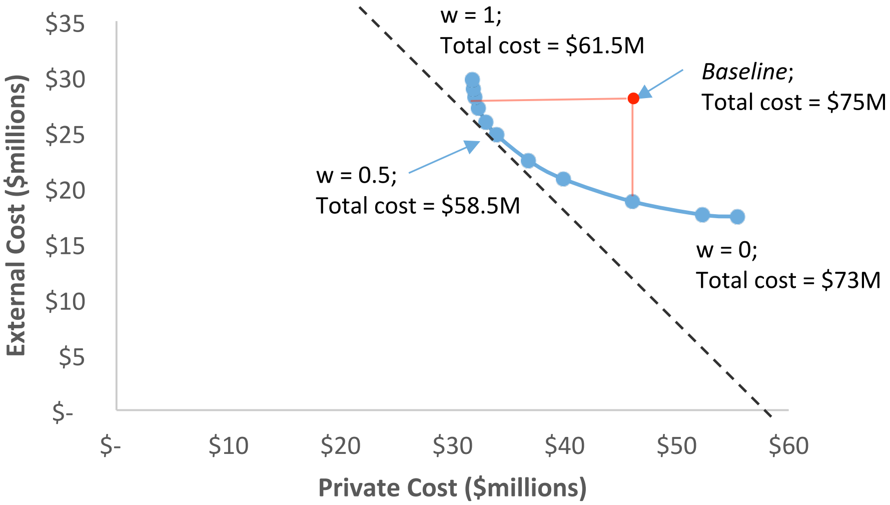


Figure 4.26: Pareto frontier for joint minimization of private and external costs. Total cost minimization where the dashed isocost line is tangent to the curve, at $w = 0.5$. The red point on the interior of the curve represents the costs of the baseline strategy. The region formed by this point and the curve defines the area in which both external and private costs can be reduced. Private costs are calculated using dynamic industrial retail prices; external costs are calculated using EASIUR MDFs.

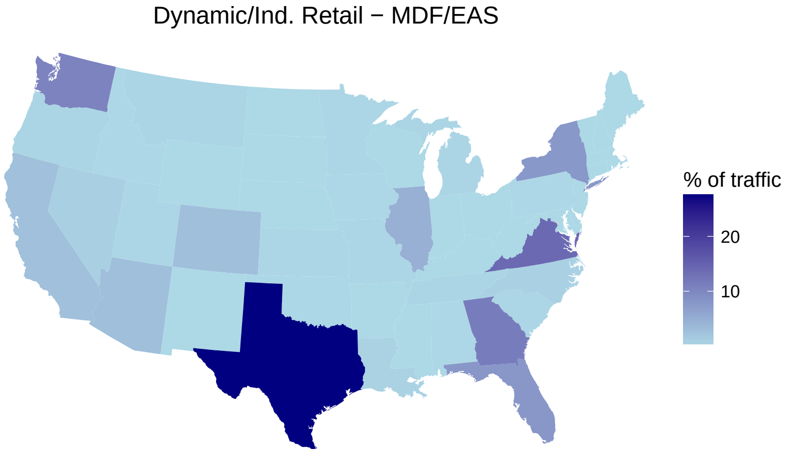


Figure 4.27: Heat map of traffic load under total cost minimization, using LMPs and EASIUR MDFs. Though only a handful of states capture the bulk of the traffic and are distinct on the map, 44 states are utilized by this strategy.

heavier use of Texas and Virginia than both scenarios, but uses California more than the private cost scenario.

4.5 Side effects of load shifting

The results thus far portray an opportunity for both public and private electricity cost savings. However, a load shifting strategy may increase other costs, potentially offsetting any power bill reductions. We now investigate the impact shifting might have on other important line items on the CDN operator's balance sheet: revenue and bandwidth costs.

4.5.1 Latency and revenue impacts

Latency may have an adverse effect on customer satisfaction. While some traffic types, such as electronic trading [256] and telepresence [257] demand low-latency service, others are much more tolerant of delay. Videoconferencing, for example, requires latency of 150ms or less, while streaming (i.e., non-interactive) video can tolerate latency of up to 5 seconds [258]. For context, the minimum fiber-optic round-trip latency from New York to San Francisco is 40 ms and from New York to London, 56 ms [259]; Verizon's enterprise networks have an average round-trip latency of around 35 ms regionally in North America, with an upper-bound guarantee of 45 ms; the company's trans-Atlantic links have latency from 70-75 ms, with a guaranteed limit of 90 ms [260]. Google's RTT from California to India is typically 300 ms [261].

One potential result of load shifting is a reduction in revenue due to increased latency, since "laggy" services are generally unacceptable to users. Revenue impacts of latency will vary greatly depending on traffic type, client, and application, but there are a few well-publicized data points from experiments by search and e-commerce firms. We should note that these sort of data points tend to come from internal experiments, are sporadically reported informally in presentations rather than published studies, and bounce around the Silicon Valley blog echo chamber, which makes tracing them back to a definitive primary source difficult. While Google and similar companies have no doubt run many interesting experiments, only a few results seem to have made it to the general public. Therefore, these numbers should be treated as anecdotal. Bing saw revenue per user declines ranging from 1.2% at a delay of 500 ms to 2.1% at 2000 ms; delays under 200 ms saw no decline in revenue [262]. Google found declines in

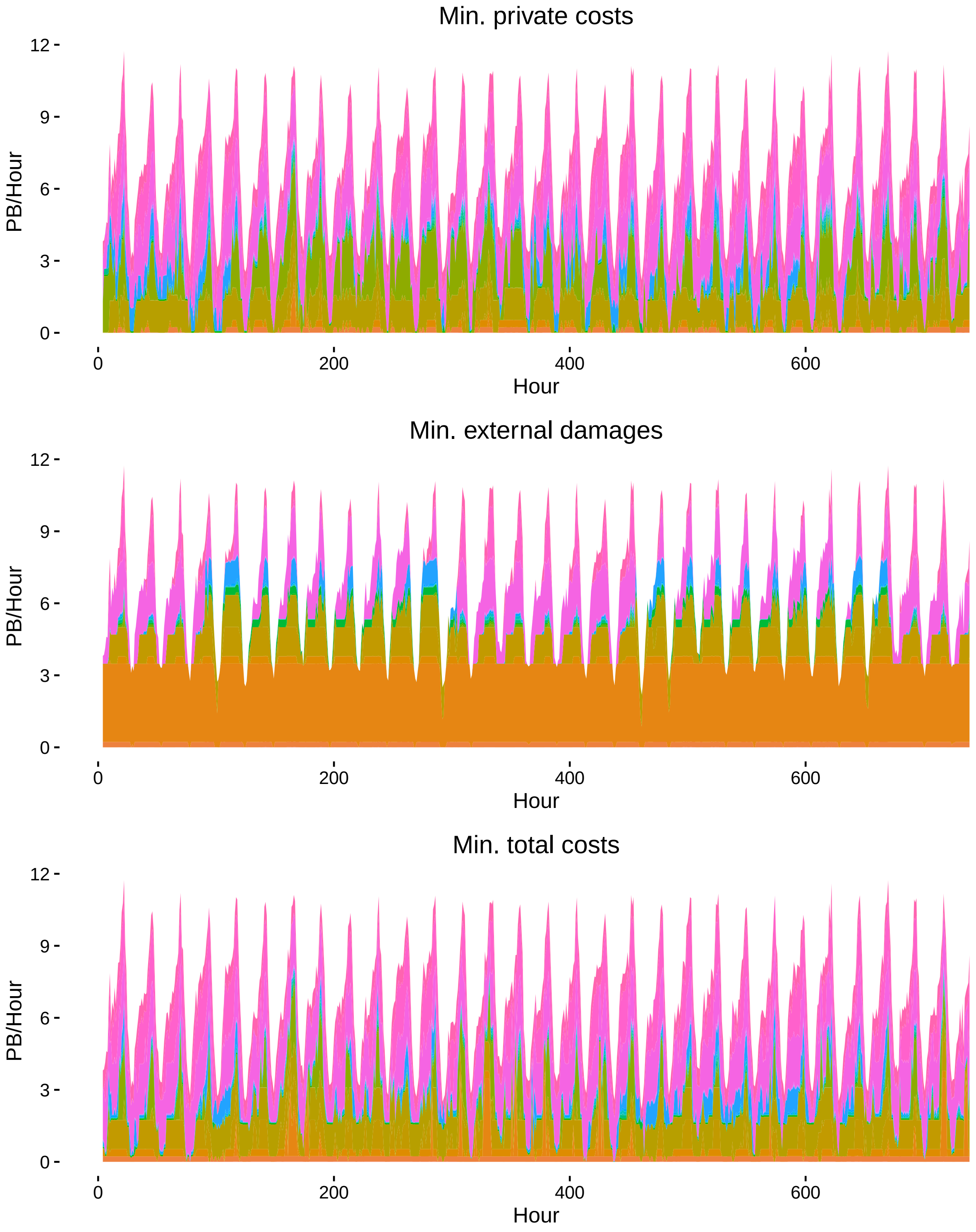


Figure 4.28: Time series of traffic loads under private (top), external (middle), and total cost (bottom) minimization strategies. Different colors represent load allocated to different data centers; heavily used data centers moving from bottom to top include California (orange), Georgia (green), New York (bright blue), Texas (purple), and Virginia (dark purple).

daily searches per user of 0.2% to 0.6% for imposed delays of between 100 and 400 ms [263]. Amazon found a sales hit of 1% for every 100 ms of latency [264]. An infographic ostensibly based on Akamai data claims that a 1-second delay in page response reduces conversions (i.e., sales transactions) by 7%, though the chart provides no background or justification for this value [265].

Clearly, latency can matter: at Google's scale, even a decline of half a percent in revenue is big. However, these numbers do not necessarily dispel the idea that an efficient CDN could add a few tens of milliseconds of latency—at least for certain types of traffic—without making load shifting infeasible. Unfortunately, we do not have any insight into the type of traffic in the Akamai data set other than that it is general "web" traffic. (Akamai removed non-web traffic, which is potentially not shiftable, prior to providing the data.) We can look at the distances in the Akamai data sample and see that in Akamai's current load-balancing algorithm, about 15% of the traffic travels farther than the distance from New York to San Francisco, although almost 3/4 of it travels a distance shorter than the New York–Chicago distance (Figure 4.29). Thus, distance is not a constraining factor for at least some CDN traffic.

Nonetheless, we now imagine that latency *does* limit the distances over which load can be shifted and investigate the effect on potential savings. So far, we have treated the entire traffic load as shiftable to any data center. We now activate a distance constraint, Equation 4.9, in the optimization problem to assess the impact on savings if traffic is limited in how far it can travel from server to client.

Imposing a universal distance limit small enough to affect intra-U.S. traffic would create an infeasible optimization problem, since U.S. nodes serve traffic from around the globe, and there are no foreign nodes in our data set to which this traffic can be routed. Therefore, we impose the distance constraint only on clients in the continental U.S., which constitute 67% of all traffic. Figure 4.30 shows the impact of distance constraint stringency on achievable private and external cost savings when either is minimized. The constraint removes flexibility to fully utilize inexpensive electricity far away, so the potential cost savings is reduced as the distance constraint becomes more stringent.

The distance constraint starts to have an increasing impact on savings below around 1,000 miles. Constraints lower than approximately 500 miles make the problem infeasible for the same reason we exempted foreign traffic: the nearest neighboring data center for some states is either too small to handle the traffic or is farther away than the allowed transport distance (particularly in the western U.S., where data centers in our model are 300-400 miles apart).

A 1,000-mile distance corresponds to a minimum RTT on the order of 15-20 ms based on propaga-

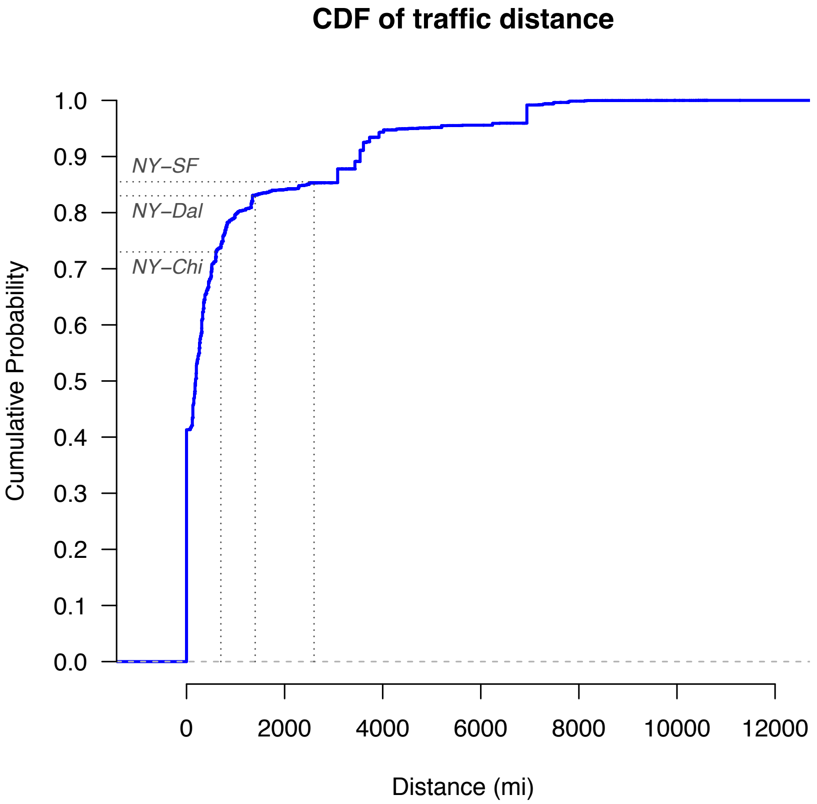


Figure 4.29: Empirical cumulative distribution function (CDF) of client-server transmission distances in Akamai sample. Gray dashed lines indicate distances from New York to Chicago, New York to Dallas, and New York to San Francisco moving left to right, respectively.

tion delay—well within the user experience thresholds defined above. Actual latencies will be higher—perhaps substantially—due to last-mile latency [259] and routing inefficiencies. Krishnan *et al.* [261] found high levels of "latency inflation" in Google's CDN due to routing inefficiencies and packet queuing; they do not attempt to solve the queuing issue, but they find that routing inefficiencies can be identified and solved. These effects may be orders of magnitude greater than the additional distance-related latency imposed by load shifting.

The relative stability of savings above 1,000 miles results from the fact that the preferred nodes in each case are distributed around the country (e.g., Washington, Texas, Illinois, Virginia, and Georgia in the private cost minimization case and California, Texas, Virginia, and Florida in the external cost minimization case; see Figures 4.20, 4.21, and 4.24). Most clients are within 1,000 miles of a node used in the unconstrained case, so increasing the allowed distance above this limit further reduces savings by only a small amount. The effect of the constraint is more prominent when external costs are minimized be-

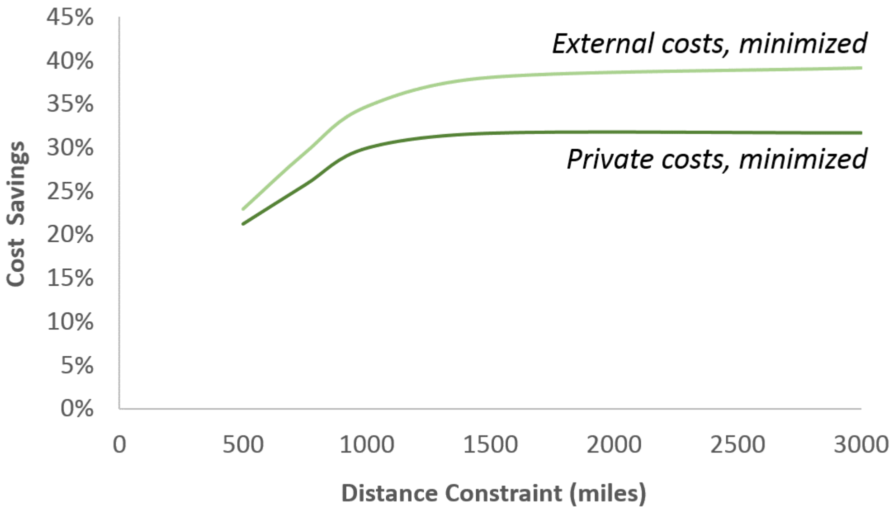


Figure 4.30: Effect of a distance constraint on private and external cost savings. Private costs are minimized using dynamic industrial retail prices; external costs are estimated using EASIUR MDFs.

cause that strategy heavily favors California, which is not centrally located and is thus made inaccessible to more clients as the constraint is made more stringent.

We note that external costs associated with minimizing private costs do not increase over this same range of distance constraints; in fact, they trend to a reduction of 5% as the constraint moves from 3K miles to 500 miles. Under external cost minimization, private costs decrease by 9% as the distance constraint decreases over its range. Thus, the distance constraint applied to private cost minimization forces use of more expensive electricity that happens to be slightly cleaner; conversely, applied to damage minimization, the constraint forces use of dirtier but slightly cheaper electricity.

If distance-based latency is a concern for the CDN operator, imposition of a distance constraint on the order of 1000 miles still leaves the vast majority of both private and external savings available. Roughly speaking, a distance constraint of this magnitude would divide the country in half, requiring two regions for serving traffic. This finding also fits with the anecdotal information, gathered from conversations with Internet industry experts, that East Coast and West Coast hubs have generally been sufficient to serve U.S. traffic, with an optional Midwest hub “if you want to pay for it.”

4.5.2 Bandwidth and operating cost impacts

We discussed the potential of load shifting to increase the 95/5 peak bandwidth costs in Section 4.2.4. We assess the impact of load shifting on the bandwidth costs two ways: calculating the break-even bandwidth price, which is the per-unit bandwidth price at which the electricity savings no longer offset the capacity charges; and imposing a bandwidth constraint on the model.

Break-even bandwidth price

To determine the break-even bandwidth price, we calculate the change in the 95th percentile of bandwidth usage at each data center under load shifting compared to the baseline, sum these changes, yielding the total additional peak bandwidth required, and then find the break-even bandwidth price as:

$$P_{\text{BE}} = \frac{\text{electricity savings}}{\text{total increase in } 95^{\text{th}} \text{ percentile bandwidth}} \tag{4.17}$$

We assume that nodes with reduced usage will be billed for less bandwidth, that bandwidth prices are uniform across nodes, and that the increase in traffic per hour required is spread uniformly across that hour.[15]

Across the range of retail price minimization scenarios, break-even bandwidth prices are approximately $1,000 per Gbps (peak) per month. The flat-rate pricing scenarios see break-even bandwidth costs of $1,100, while the dynamic scenarios see break-even costs of $900. This difference results from the more aggressive shifting undertaken by the dynamic strategies, which use more nodes than the flat-rate strategies, thereby creating a large increase in bandwidth use to achieve a smaller increase in electricity cost savings. Private cost minimization increases 95/5 peak bandwidth use by 4%-6%.

If additional bandwidth can be obtained at a cost less than the break-even prices, then load shifting remains viable. Actual bandwidth costs can range from $1 to over $10 per peak Mbps per month [266, 267].[16] Our rough break-even price of $1,000/Gbps per month translates into $1/Mbps per month, which is the bottom end of the bandwidth price range. While peering and bilateral agreements might allow a large CDN to reduce these rates somewhat, these calculations indicate that cost-minimization without bandwidth constraints is only cost-effective if bandwidth can be obtained very cheaply.

[15] We are limited to hourly resolution, so our bandwidth is measured in the unconventional metric of Gb served per hour. We can convert this metric to an average Gb per second (Gbps) rate over the hour.

[16] In addition to the sources cited, these figures are informed by conversations with industry experts.

However, several aspects of the problem may mitigate this issue. First, this calculation is highly sensitive to the energy conversion factor (i.e., kWh/GB). We have assumed very efficient data centers; as noted above, the estimates for the energy conversion factor range over several orders of magnitude, so assuming a less efficient data center would raise the break-even prices by increasing numerator in Equation 4.17. Second, our optimization model is "greedy" in the sense that it seeks to extract all possible savings from price differentials. A more balanced model could include bandwidth costs and prevent the optimizer from shifting load in ways that dramatically increase bandwidth; we explore the extreme version of such a model in the next section. Third, bandwidth prices are continuing to decrease, so the $1/Mbps monthly price may be less restrictive in the future.

Under external damage minimization, break-even bandwidth costs range from $300 to $1,500, depending on the damage model used. EASIUR, in particular, drives up peak bandwidth by 12–16%, whereas the increase under AP2, at 2–5%, is more in line with the private minimization strategies. The external break-even price represents the maximum subsidy that should be paid to a damage-minimizing CDN operator to allow them to recover increased bandwidth costs; however, such a subsidy would also need to cover any increase in electricity costs, so the external break even bandwidth price should be lower.

Bandwidth constraint

Employing the constraint in Equation 4.10 ensures that the baseline bandwidth bill is not exceeded. Using the dynamic industrial retail pricing scenario, enforcing the existing 95/5 constraint reduces private cost savings from 32% to 25%. That is, about 80% of the potential electricity cost savings under load shifting can still be achieved without increasing bandwidth charges. At the same time, the effect on damages is an increase from 4% to 7% over the baseline. Thus, there is ample headroom even under the existing bandwidth cap to shift load and capture most of the savings. Obviously, the CDN operator could relax the bandwidth constraints to reduce electricity costs further.

4.5.3 Summary of side effects from load shifting

Ultimately, latency-tolerant traffic and abundant bandwidth increase the viability of load shifting. At the same time, load shifting appears viable even if latency and bandwidth constraints are present. Finally, this analysis, to some extent, focuses on the impacts of the extreme "edge cases." Together with the tradeoffs shown in Figure 4.26, these findings indicate that there should be a strategy in which both private and external costs can be reduced and the additional latency and bandwidth impacts can be mitigated by being slightly less aggressive in shifting load.

4.6 Investment options for load shifting and distributed renewable generation

To this point, our analysis has determined the potential savings available under load shifting using existing CDN infrastructure. We now briefly turn to new investment and evaluate two strategies. First, we assume that load shifting is undertaken and assess the additional savings available by investing in new node capacity. Second, we compare the ability of load shifting to reduce electricity costs to that of distributed renewable generation.

4.6.1 Investing in capacity expansion: shadow prices of node capacities

We observed from the results and heat maps above that, while there are clear differences in the load solutions used by the different cost minimization strategies, in practice most of the traffic is served by subsets of the same few, high-capacity nodes. The capacities of most of the nodes are too small to fully leverage price and damage differentials in these states (Figure 4.3). If a smaller node tends to have low electricity costs, it might be worthwhile to invest in capacity expansion at this node. Here, we examine the shadow prices on the node capacity constraints (Equation 4.7) to determine if these limits are preventing the optimization from utilizing a location with a particularly compelling cost structure.

Before doing so, Figures 4.10 and 4.16[17] provide some insight into what we should expect. Based on LMPs, we would expect the upper Midwest nodes, in addition to Utah and Oregon, to have the largest shadow prices.[18] Looking at industrial retail prices, Texas, New York, the Southeast, and the Pacific Northwest look like candidates for expansion. Turning to damages, California, Arizona, and New Mexico have the lowest damages and should have the highest shadow prices. Note that these expectations are based on the average cost differentials, so the hourly dynamics could make other nodes appealing for expansion.

[17] See also Figure A.36 in the Appendix.

[18] As this is a minimization, the shadow prices are actually negative—i.e., they provide the decrease in cost available by relaxing the constraint. Thus, by *largest*, we mean largest in magnitude.

At the same time, there is evidence that the existing capacity constraints are not likely to be large inhibitors to cost savings. First, in the current network topology, the subset of large nodes includes both clean states (California) and cheap states (Texas), so relaxing capacity constraints—while providing the opportunity to achieve additional savings—is unlikely to result in radically different load maps. Second, the relatively large amount of headroom available in the network indicates that a high degree of slack or flexibility to shift load around already exists in the system. Consequently, we expect the benefits of relaxing node capacity constraints to be relatively small.

When minimizing private costs, the shadow price on a node capacity constraint represents the decrease in electricity cost made available by being able deliver one additional TB per hour through that node. Under industrial retail pricing, the shadow prices at five nodes exceed $900: Washington approaches $2,000, while Kentucky, Montana, Louisiana, and Oklahoma are near $1,000. Under commercial retail pricing, the top fives states are Idaho, Oklahoma, Texas, Washington, and Virginia, with shadow prices from $1,000 to $1,300.

Under external cost minimization, the shadow price represents the value of damages that can be avoided by adding the capacity to deliver an additional TB per hour at a particular node. As we expected, under marginal damage models, California, Arizona, and New Mexico are far and away the best candidates for expansion, with shadow prices in excess of $1,300. The EASIUR model places Florida and Colorado next, at around $500, while AP2 favors the Northwest, with prices around $800. Using average damages, EASIUR again places the highest value on California ($1,300), Arizona and New Mexico ($600), and then Florida ($500) and New England ($400), while AP2 values New England capacity on par with California, both at $1,000, followed by Arizona and New Mexico at $700.

In sum, $1,000 per TB of hourly traffic capacity seems to be a good rough estimate for the value of adding data center capacity, regardless of whether the CDN operator is investing to further reduce electricity bills or the public is paying to reduce damages. If the cost of capacity expansion is lower than the shadow price, then expansion may be cost effective. The cost of adding this capacity would include fixed costs such as hardware and labor and any variable costs not included in the optimization model. We suspect that, in many circumstances, the shadow price is lower than the actual cost of expansion, although the value really depends on the specifics of the expansion contemplated.

4.6.2 Investing in distributed generation: average electricity prices

We compare the load shifting strategy with another potential cost-reduction approach: building renewable distributed generation. Figure 4.31 compares the average cost of electricity under the baseline and load shifting strategies with the levelized cost of electricity (LCOE) for wind, solar PV, and hydrogen fuel cell distributed generation technologies. The LCOE estimates come from LBNL [268, 269], the Energy Information Administration (EIA) [270], and Lazard [271]. The LBNL estimates are from reviews of power purchase agreements and include subsidies, so the figure reflects an upward adjustment on solar of 30% and on wind of $20/MWh, to provide a rough estimate of unsubsidized LCOE. These adjustments reflect federal incentives identified in the reports; state incentives may require further adjustments. The EIA estimates are model projections of unsubsidized LCOE in 2020. The Lazard estimates are unsubsidized and include utility-scale plants as well as commercial and industrial rooftop solar, but we exclude residential rooftop PV. Note that Lazard's utility-scale solar estimate ranges from $60 - $86, while rooftop solar ranges from $126 - $177. Solar PV and wind are assumed to have zero marginal damages. Load shifting figures are shown as both private and total costs, with high and low values reflecting commercial and industrial retail prices, respectively. Total costs are private costs plus external damages.

The figure shows that wind and solar generation are generally less costly than load shifting, even when only private costs are considered. Note that these LCOEs reflect *distributed* generation costs—i.e., a scenario where the data center either builds its own plant (see Apple's North Carolina facility [14]) or executes a power purchase agreement with such a plant; these estimates may not reflect the cost of procuring energy through the data center's regular electricity supplier via a "green power option." Furthermore, the economics of distributed generation are site- and situation-specific, and the overlapping range of estimates indicates that there may be circumstances where private cost minimization is preferred to distributed generation.

Publicity, ethics, stewardship, and other strategic benefits may cause the CDN operator to consider these options as more than cost reduction tools, and once a CDN operator has decided to take a "green" perspective, the figures favor distributed generation over lead shifting more heavily. Additionally, the load shifting results represent fairly aggressive electricity cost minimization, so bandwidth and latency priorities may increase these costs somewhat. Given the location-specific cost variability in distributed generation, the broad ranges reported here, and the uncertainty in the load shifting analysis, it is difficult

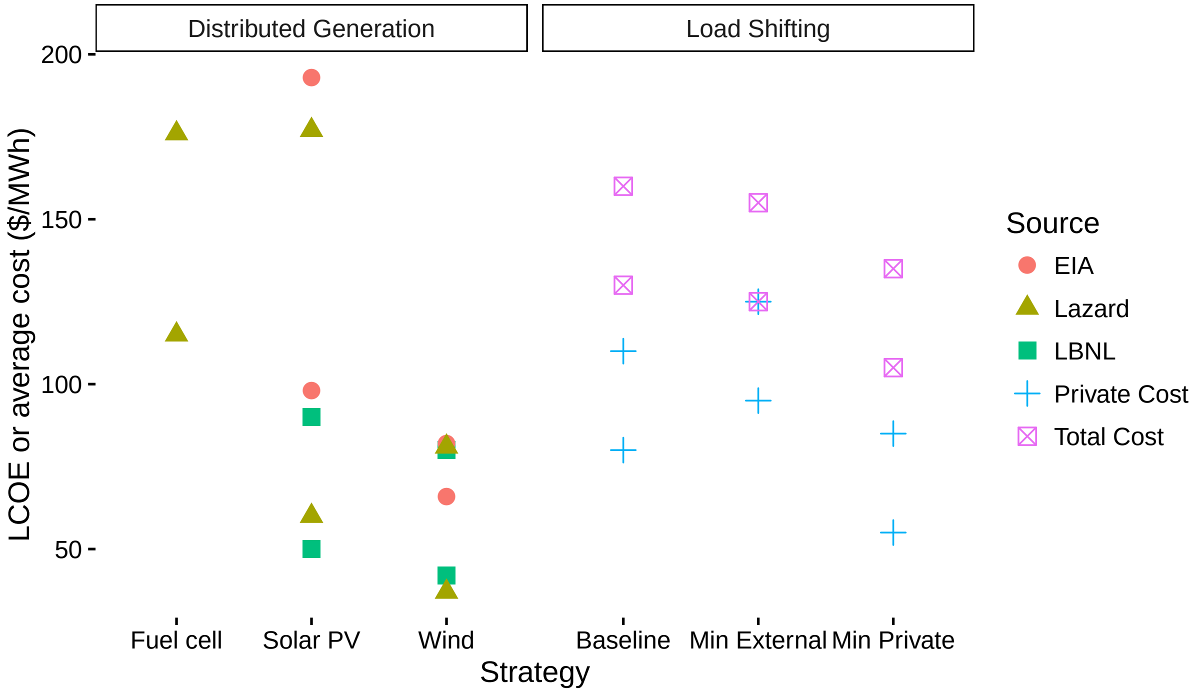


Figure 4.31: Average electricity cost comparison between load shifting and distributed generation strategies. See text for sources and notes.

to dismiss load shifting completely. Additionally, while the evidence suggests that load shifting, in the average case, is more expensive than distributed generation, it does have some cost advantages: it still reduces electricity costs vs. the baseline, it can be undertaken with minimal capital outlay, and it carries no long-term commitment and can easily be reversed.

4.7 Synthesis of results

Load shifting is an interesting idea. The analysis here suggests that potential retail electricity cost savings of 23% and 32% are possible under the private cost minimization strategy, depending on the electricity price structure faced by the CDN operator. Similarly, avoided damages range from 29% to 39% under the damage minimizing strategy. There is a broad range of weightings in which joint optimization substantially reduces both external and private costs simultaneously, suggesting common ground that could benefit both CDN operators and the public.

However, several barriers decrease the appeal of load shifting. Dynamic pricing is not a common rate structure for data centers. It may not be available in all areas, and there may be transaction costs associated with moving to such a rate structure where it is available. Further complicating this matter is the fact that CDN nodes are typically housed in colocation centers, the operators of which would act as an intermediary between the CDN owner and the electric utility. Colocation centers typically pass electricity costs through to the CDN owner; thus load shifting would require that the colocation provider opt for RTP and make that dynamic pricing information transparently available to its clients in real time. There may also be regulatory issues dictating how these costs can be passed through.

The analysis of bandwidth costs suggests that the additional peak bandwidth demanded by the load shifting strategies examined here might, if unchecked, erode electricity savings. However, a less greedy arbitrage strategy in which the optimizer does not allow bandwidth costs to increase captures most (80%) of the available energy cost savings. Furthermore, we use a comparatively low conversion factor to transform traffic to energy consumption at the node, implying a very efficient data center. Were we to use one of the higher estimates for data center energy use, the bandwidth costs would become less significant compared to electricity savings. Finally, bandwidth is continuing to get cheaper and may thus become less of a barrier in the future.

This analysis suffers somewhat from uncertainty, particularly in the energy conversion factor at the data center and the various adders used to translate LMPs to retail prices. As constructed, the results scale linearly with the energy conversion factor, so it is straightforward to assess the impact of different values of energy consumption per unit traffic. However, data center nodes may not, in fact, be power proportional, in which case a nonlinear model would be more accurate.

The nature of price composition using regional averages (see Section 4.3.2) adds uncertainty to our price estimates and hence to the savings from load shifting obtained by our model. Importantly, the bulk of achievable savings can be obtained by exploiting regional differences in price (i.e., without leveraging hourly variability), and our flat-rate scenarios *do not* depend on this construction of retail prices; rather, they use rates reported by the EIA. Thus, only the additional savings obtained by leveraging RTP are affected by the uncertainty in the adders.

Additional uncertainties lie in the proportion of traffic that can actually be shifted and bandwidth costs. Thus, these results provide only an indication of possible savings. A CDN operator contemplating arbitrage would need to revise this model with power consumption, electricity pricing, traffic load, network topology, bandwidth cost, and traffic type data specific to his network to get an accurate picture of expected savings.

We can do a (very) rough check to see if the savings estimated are reasonable for Akamai. The model estimated on the order of \$12–\$17 million in energy cost savings across the U.S. nodes under retail pricing. Qureshi *et al.* [204] estimated a total Akamai power bill of \$10 million, associated with 40K servers, seven years ago. Akamai now has 216K servers [272]. If energy use scales linearly with server growth, Akamai's power bill is now on the order of \$50-\$60 million. Akamai's servers are likely to be more energy efficient now than they were in 2009, so if we use the low end of this range, our model suggests potential electricity bill savings of 24% to 34%. These savings are substantial, and—though realized savings are likely to be lower for the reasons mentioned—they are within the range of results reported by Qureshi *et al.* [204]. In the context of the corporate balance sheet, these savings amount to 2% of Akamai's annual revenue and 11% of operating income [273].

The average electricity cost estimates for load shifting overlap with LCOE estimates for distributed renewable generation, meaning that the minimum-cost option is site-specific. When external costs are considered, wind and solar will almost certainly yield lower total costs.

Despite the caveats discussed above, electricity cost minimization via load shifting shows some promise for both private and public electricity cost reductions. Data center networks are perhaps the only industrial or commercial entity that can quickly shift load geographically, and therefore this sort of load shifting could act as a useful complement to the temporal peak-shaving and demand response programs implemented in other industries.

In order for load shifting to work, however, real-time pricing must be broadly adopted by the colocation centers that host CDN nodes. Furthermore, the savings opportunities identified here require that RTP be closely indexed to wholesale LMPs; other variable rates with lower volatility may not achieve the same level of savings. Finally, the colocation center must continue to pass electricity costs through to their clients; if the colocation host attempts to profit from the arbitrage opportunity (e.g., subscribing to a variable rate but charging clients a flat rate), then it is possible that the colocation center will fall under regulations that apply to public utilities and utility resellers.

Table 4.9 summarizes the effect of different industry trends on the size of the savings achievable through the load shifting strategy. It is probably bad news for load shifting that the three most important trends in reducing data center energy impacts—increasing energy efficiency, increasing utilization, and green power options[19]—reduce its benefit. Nonetheless, a strategy that would reduce electricity costs by

[19] See Chapter 2.

1/4 to 1/3 using existing infrastructure is worth investigating. The most logical path to implementation would be for electricity cost differentials to be added as a component in the existing routing algorithms employed by CDNs. In this way, the electricity cost optimization could provide an additional degree of flexibility without overwhelming other important factors, such as bandwidth costs and traffic priorities.

Table 4.9: Effect of current trends on viability of load shifting.

Trend	Benefit to load shifting	Rationale
Decreasing bandwidth costs	Positive	Load shifting, when unconstrained, can markedly increase peak bandwidth usage, eating into electricity cost savings. Cheaper bandwidth reduces this impact.
Consolidation of traffic in the cloud	Positive	Load shifting is likely only worthwhile for larger networks of data centers that handle lots of traffic
Electricity wholesale market restructuring	Positive	Market restructuring and consolidation of regional wholesale markets with real-time prices creates the basis for the arbitrage strategy.
Dynamic retail electricity rate structures	Positive	Access to dynamic retail prices makes load shifting viable, with higher-frequency price changes being better.
Increasing proportion of on-demand video streaming traffic	Negative (?)	On one hand, on-demand streaming is latency tolerant, which would be positive; however, high bandwidth requirements of streaming mean that it may be constrained to edge servers to prevent bottlenecks, which would preclude load shifting.
Increasing data center energy efficiency	Negative	The more energy consumed to serve a unit of traffic, the greater the electricity savings from shifting the load.
Increasing utilization/virtualization	Negative	High utilization reduces available slack to shift load.
Decreasing costs of distributed renewable generation	Negative	Distributed renewable generation is a competing approach; the cheaper it is to source green power, the more likely it is to be preferred to load shifting.

We should also note that this analysis reflects opportunities available for one CDN, which we assume is a price-taker and thus has no marginal effect on electricity prices. What if, on the other hand, all data center networks adopted load shifting? Is it conceivable that this strategy would affect electricity markets? We leave a detailed assessment of this question for future work, but our sense is that such a consequence is unlikely. Data centers consume only 2% of U.S. electricity, and only a fraction of these data centers are CDNs or other networks that could shift load—although, as we noted earlier, CDN electricity consumption is likely growing faster than consumption in other types of data centers. Furthermore, not all traffic is shiftable: large-scale streaming video often needs to be served from the edge of the network, as previously discussed, and applications dependent on maintaining state data also impose complications on shifting. Finally, while large data centers may be non-marginal electricity consumers at the local level, these facilities are typically accompanied by the grid infrastructure necessary to support their demand, which should mitigate large deviations from the regional grid LMP.

One potential use of load shifting which we have not explored is as an ancillary service provider for the electricity grid. The demand response and peak-shaving applications are obvious, but the unique loads of data centers can perhaps support other grid requirements, particularly when their large battery banks are considered. A thorough analysis of the demand response opportunity is left for future work, but we can provide a rough estimate of peak-shaving potential from our model. The peak data center electricity demand by the California node of our modeled CDN is 25 MW, which alone is not enough to make much of a difference for peak-shaving; however, we note that it is on the scale of some of the smaller gas power plants designated in California as "emergency peakers." Furthermore, the modeled CDN is only one of many CDNs with nodes in California, so it is conceivable that load shifting could play a role in demand response events.

Because minimizing the power bill may increase external damages, if data center operators do begin to utilize price-responsive load shifting, policymakers should then provide a mechanism to ensure that externalities are also factored into the equation. The analysis here suggests that such an incentive could target a broad zone in the joint optimization space where substantial public and private savings can be achieved simultaneously.

Chapter 5

Conclusions: Policy Implications

This dissertation examines three different aspects of data center and ICT energy use: individual data center energy metrics, the indirect energy impacts of ICT service deployment, and strategies for energy cost optimization across a network of data centers. Here, we highlight opportunities for guidance through policymaking in each of these areas.

5.1 Metrics

The assessment of existing and proposed metrics revealed crucial gaps between current assessment and reporting and what is needed to accurately and comprehensively measure data center energy performance. PUE is, at best, an incomplete metric, incorporating facility overhead but saying nothing about IT equipment efficiency, utilization, or power sourcing. At worst, it is a perverse incentive: in some cases, steps taken to improve PUE are undesirable from an overall energy standpoint. To address these gaps, government leadership is needed in three key areas.

Metrics development. The analysis made it clear that PUE is, by itself, insufficient to properly incentivize data center energy efficiency. Focusing solely on reducing PUE shortchanges improvements that can be made in equipment efficiency, utilization, and power sourcing. By encouraging the adoption of new metrics through cooperation with industry working groups and sponsoring research on new metrics, the government can help overcome the inertia of a PUE-focused industry.

Standards development. Data center energy efficiency standards are still in their infancy, and a range of governmental and nonprofit organizations still heavily reference PUE in these standards. Updating programs like the ENERGY STAR score for data centers with a relevant and comprehensive suite of metrics would help the industry move towards a more complete view of data center energy performance. As discussed in Chapter 2, ENERGY STAR has a separate program for rating IT equipment. Making the latter program an input to the data center rating would be a logical first step, particularly since improving the efficiency of IT equipment in an existing facility is likely to make PUE *worse*.

Data collection and availability. The government can improve the lack of measured data on computing facilities through several means. First, it can sponsor evaluation and data collection programs like those carried out at LBNL and used in this study. Second, it can encourage industry partners to collect and publish non-proprietary data. Finally, government organizations operate a large number of data centers and collect their own metrics. Raw data collected from standardized reporting programs like the Federal Energy Management Program and through initiatives like the Federal Data Center Consolidation Initiatives should be made available for analysis by the research community.

5.2 Guiding wide-scale ICT deployment

The review of indirect effects showed that, while there is a pervasive—and not unwarranted—optimism about the potential of ICT to improve societal energy efficiency, the realized energy effect may be positive or negative, depending on how ICT is deployed. Some care should be taken in helping guide ICT deployment toward scenarios where it acts as a damper, rather than an amplifier, on energy consumption. The government and its policymakers have several important roles in this guidance.

Envision future scenarios. There are many divergent ways ICT could evolve. Taking autonomous vehicles as an example, one future would see increased migration to cities as individuals take advantage of ubiquitous urban fleets of public or shared driverless vehicles to alleviate the need for personally-owned vehicles (POVs). An alternate future would see an exacerbation of suburban sprawl, as driverless POVs increase the utility of commuting by allowing car occupants to accomplish other tasks (e.g., telework, entertainment, or napping). Through strategies like scenario analysis and visioning exercises, the government can engage stakeholders to characterize these alternate pathways.

Support data gathering and research. Government modeling and data collection, through entities like the EIA and the National Laboratories, provide some of the best sources for assessing impacts and trends. More quickly updating models like the National Energy Modeling System to account for developments in ICT infrastructure and services would help ensure that these resources stay relevant and aid our understanding of the effects of ICT deployment.

Incentivize socially-beneficial choices. After fostering a better understanding of possible future scenarios, policymakers can then act to support choices that lead to desired or socially-beneficial outcomes. Using the autonomous vehicle example, for instance, providing grants or loans for urban redevelopment would support the first outcome, while increasing highway funding would support the second outcome.

5.3 Cost-responsive routing

The buik of the analytical work of this dissertation focused on the load shifting study, which showed that routing to minimize private costs could reduce them by 25%–33%, that routing to minimize external damages could reduce them by 30%–40%, and that there is a solution space that allows simultaneous savings in both cost metrics. Furthermore, the study showed that a large majority of potential cost savings could be achieved even when limiting transport distances and peak bandwidth consumption to existing levels.

While there remain some questions about the appeal of this strategy moving forward, policymakers have the opportunity to remove some of the barriers to the adoption of this strategy.

Expand dynamic electricity rate schedules. Consumer exposure to dynamic electricity pricing is not widespread, despite traditional economic consensus that it is more efficient than a flat-rate pricing scheme. As electricity is a public good, the federal and state governments generally have a mandate to regulate the sector, so a policy push for dynamic pricing could have a positive impact on the growth of these rate schedules.

Support electricity market improvements. Large portions of the country do not participate in a regional electricity market, making exposure to LMPs somewhat more difficult. Full nationwide coverage by regional wholesale electricity markets would best support the dynamic load shifting strategy contemplated here.

Align electricity data reporting. The electricity sector benefits from abundant data collected and disseminated by government and pseudo-government entities. Between EIA, the EPA, FERC, NERC, and the individual ISOs, data on generation, demand, prices, and emissions are generally available. Unfortunately, harmonizing these data is a frustratingly difficult task, primarily because of the boundary mismatches among the regions used in these different reports. NERC, eGRID, ISO, and BA boundaries are somewhat reflective of the legacy electricity grid, and the consolidation of markets and merging of data has muddied these boundaries. A regional overlay with consistent boundaries, along with an explicit mapping between the same entities in different data sets, would make analysis of generation, demand, emissions, and other aspects of the electricity system more straightforward and less prone to error.

Incentivize awareness of social electricity costs. The analysis of load shifting showed that if CDN operators do begin price-responsive load shifting, the strategies they adopt to reduce their own costs may leave significant social benefits related to avoided damages on the table. Policymakers are already implementing electricity-related cost incentives for data centers; for example, North Carolina, which has developed a hub of warehouse-scale data centers, exempts qualifying data centers from electricity taxes [274]. These sorts of incentives should include both the public and private perspective on costs.

5.4 Looking forward

The continued development and deployment of ICT is inevitable and, in many ways, incredibly exciting. Despite regular periodic grumblings about "simpler times," and the key privacy, security, and equity challenges brought about by the rapid growth of this industry, by many objective measures, ICT innovation is transforming our society and economy in extremely beneficial ways. ICT has fostered increased transparency; provided greater information access, more efficient commerce, and additional outlets for creativity; lowered barriers to entry in many industries; helped us maintain personal connections across time and space; made it easier to gain exposure to different cultures and heterogeneous groups; and helped increase our understanding of the world through scientific exploration and analysis.

The energy and environmental legacy of this hugely important ICT sector is yet to be determined. ICT certainly has the great potential to further increase its own energy efficiency and to reduce energy consumption and impacts throughout society and the economy. However, such an outcome is by no means assured, and may not come about if we allow ICT to develop in an *ad hoc* manner outside of a policy framework. Such a framework need not constrain the development of ICT. Done correctly, the sorts of government actions outlined here—development of scenarios, metrics, standards, and data collection—will only help ICT serve as a "force multiplier" for achieving our societal, economic, and environmental goals. Incentives which help align technology deployment with social good, such as those that internalize externalities or nudge the trajectory toward a desired end state, can be productive without infringing on the independence, excitement, flexibility, and entrepreneurial flair that have come to characterize the ICT industry.

References

[1] N. Horner, I. Azevedo, and D. Hounshell, "Effects of government incentives on wind innovation in the United States," *Environmental Research Letters*, vol. 8, no. 4, p. 044 032, Dec. 2013, ISSN: 1748-9326. DOI: 10.1088/1748-9326/8/4/044032. [Online]. Available: http://iopscience.iop.org/article/10.1088/1748-9326/8/4/044032/.

[2] LBNL, *Definition of Data Center*, 2013. [Online]. Available: http://energy.lbl.gov/ea/mills/HT/dctraining/definitions.html.

[3] Gartner, *Gartner IT Glossary - Data Center*, 2013. [Online]. Available: http://www.gartner.com/it-glossary/data-center/.

[4] W. P. Turner, J. H. Seader, V. Renaud, and K. G. Brill, "Tier Classifications Define Site Infrastructure Performance," Uptime Institute, White Paper, 2008, p. 19. [Online]. Available: http://www.de-graft.com/wp-content/uploads/2012/09/Tier-Classifications-Define-Site-Infrastructure.pdf.

[5] C7 Data Centers, *C7 tier classifications*, 2012. [Online]. Available: http://www.c7dc.com/articles/tier-classifications/.

[6] M. K. Patterson, *The Green Grid EPA Data Center Assessment*, 2010. [Online]. Available: http://www.thegreengrid.org/~/media/TechForumPresentations2010/EPA_Data_Center_Assessment_Report.pdf?lang=en.

[7] L. A. Barroso and U. Hölzle, "The datacenter as a computer: an introduction to the design of warehouse-scale machines," *Synthesis Lectures on Computer Architecture*, vol. 4, no. 1, pp. 1–108, Jan. 2009, ISSN: 1935-3235, 1935-3243. DOI: 10.2200/S00193ED1V01Y200905CAC006. [Online]. Available: http://www.morganclaypool.com/doi/abs/10.2200/S00193ED1V01Y200905CAC006.

[8] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, and W. Lintner, "United States Data Center Energy Usage Report," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-1005775, Jun. 2016. [Online]. Available: http://eetd.lbl.gov/sites/all/files/lbnl-1005775_v2.pdf.

[9] R. Miller, *Microsoft Reveals its Specialty Servers, Racks*, Apr. 2011. [Online]. Available: http://www.datacenterknowledge.com/archives/2011/04/25/microsoft-reveals-its-specialty-servers-racks/.

[10] CES, "Free cooling concepts for data centers," White Paper CESG-DC-WP-2A, Jul. 2014. [Online]. Available: http://www.findwhitepapers.com/force-download.php?id=39749.

[11] D. Tuite, "400-V DC distribution in the data center gets real," *Electronics Design*, Feb. 2014. [Online]. Available: http://electronicdesign.com/power/400-v-dc-distribution-data-center-gets-real-0.

[12] C. Garling, "AC/DC battle returns to rock data-center world," *WIRED*, Dec. 2011. [Online]. Available: http://www.wired.com/2011/12/ac-dc-power-data-center/.

[13] M. Murrill and B. J. Sonnenberg, "Evaluating the Opportunity for DC Power in the Data Center," Emerson Network Power, White Paper, 2010. [Online]. Available: http://megaglobalsolution.com/files/liebert/OTHER%20TYPE/Netsure%20ITM%2048V%20DC-UPS,70-280KVA/124W-DCDATA-web.pdf.

[14] R. Miller, *Apple: iDataCenter Power Will Be 100% Green*, May 2012. [Online]. Available: http://www.datacenterknowledge.com/archives/2012/05/17/apple-idatacenter-power-will-be-100-green/.

[15] IBM, *What is big data?* Dec. 2013. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

[16] J. Markoff and S. Hansell, "Hiding in plain sight, google seeks more power," *The New York Times*, Jun. 2006, ISSN: 0362-4331. [Online]. Available: http://www.nytimes.com/2006/06/14/technology/14search.html.

[17] J. Glanz, “Data centers waste vast amounts of energy, belying industry image,” *The New York Times*, Sep. 2012, ISSN: 0362-4331. [Online]. Available: http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html.

[18] J. G. Koomey, “Rebuttal to testimony on ’Kyoto and the internet: The energy implications of the digital economy’,” *Lawrence Berkeley National Laboratory*, 2000. [Online]. Available: http://escholarship.org/uc/item/2nf809r0.pdf.

[19] J. G. Koomey, C. Calwell, S. Laitner, J. Thornton, R. E. Brown, J. H. Eto, C. Webber, and C. Cullicott, “Sorry, wrong number: the use and misuse of numerical facts in analysis and media reporting of energy issues,” *Annual Review of Energy and the Environment*, vol. 27, no. 1, pp. 119–158, Nov. 2002, ISSN: 1056-3466. DOI: 10.1146/annurev.energy.27.122001.083458. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.energy.27.122001.083458.

[20] J. G. Koomey, H. Chong, W. Loh, B. Nordman, and M. Blazek, “Network electricity use associated with wireless personal digital assistants,” *Journal of infrastructure systems*, vol. 10, no. 3, pp. 131–137, 2004. [Online]. Available: http://ascelibrary.org/doi/abs/10.1061/(ASCE)1076-0342(2004)10%3A3(130).

[21] J. Aslan, K. Mayers, J. G. Koomey, and C. France, “Electricity intensity of Internet data transmission: Untangling the estimates,” *Journal of Industrial Ecology*, vol. Submitted, 2016.

[22] J. G. Koomey, “Worldwide electricity used in data centers,” *Environmental Research Letters*, vol. 3, no. 3, p. 034 008, Sep. 2008. [Online]. Available: http://iopscience.iop.org/1748-9326/3/3/034008.

[23] J. G. Koomey, “Growth in data center electricity use 2005 to 2010,” *Oakland, CA: Analytics Press. August*, vol. 1, p. 2010, 2011. [Online]. Available: http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2011finalversion.pdf.

[24] N. Horner and I. Azevedo, “Power usage effectiveness in data centers: overloaded and underachieving,” *The Electricity Journal*, vol. 29, no. 4, pp. 61–69, May 2016, ISSN: 10406190. DOI: 10.1016/j.tej.2016.04.011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1040619016300446.

[25] D. Anderson, T. Cader, T. Darby, N. Gruendler, R. Hariharan, A. Holler, C. Lindberg, and C. Long, “A Framework for Data Center Energy Productivity,” The Green Grid, White Paper WP-13, 2008. [Online]. Available: http://lpis.csd.auth.gr/smartihu/storage/GreenGridProxies.pdf.

[26] eBay, *Digital Service Efficiency*, Corporate website, 2013. [Online]. Available: http://tech.ebay.com/dashboard.

[27] J. Whitney, J. Taylor, and C. Kral, “Salesforce.com and the environment: reducing carbon emissions in the cloud,” WSP Environment and Energy, Tech. Rep., 2011. [Online]. Available: https://www.salesforce.com/assets/pdf/misc/WP_WSP_Salesforce_Environment.pdf.

[28] E. Masanet, A. Shehabi, and J. G. Koomey, “Characteristics of low-carbon data centres,” *Nature Climate Change*, vol. 3, pp. 627–630, Jul. 2013. DOI: 10.1038/NCLIMATE1786. [Online]. Available: http://www.nature.com/nclimate/journal/v3/n7/abs/nclimate1786.html.

[29] J. R. Stanley, K. G. Brill, and J. G. Koomey, “Four Metrics Define Data Center ’Greenness’,” Uptime Institute, Santa Fe, NM, White Paper TUI3009F, 2007. [Online]. Available: http://www.dcxdc.ru/files%5C4ede4eff-13b0-49d9-b4da-b0406bfc190e.pdf.

[30] J. Hamilton, *PUE and Total Power Usage Efficiency (tPUE)*, Blog, Jun. 2009. [Online]. Available: http://perspectives.mvdirona.com/2009/06/15/PUEAndTotalPowerUsageEfficiencyTPUE.aspx.

[31] M. K. Patterson, S. W. Poole, C.-H. Hsu, D. Maxwell, W. Tschudi, H. Coles, D. J. Martinez, and N. Bates, “TUE, a new energy-efficiency metric applied at ORNL’s Jaguar,” in *Supercomputing*, Springer, 2013, pp. 372–382. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-38750-0_28.

[32] S. Rivoire, M. A. Shah, P. Ranganathan, and C. Kozyrakis, “JouleSort: a balanced energy-efficiency benchmark,” in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, ACM, 2007, pp. 365–376. [Online]. Available: http://dl.acm.org/citation.cfm?id=1247522.

[33] R. H. Katz, D. E. Culler, S. Sanders, S. Alspaugh, Y. Chen, S. Dawson-Haggerty, P. Dutta, M. He, X. Jiang, L. Keys, A. Krioukov, K. Lutz, J. Ortiz, P. Mohan, E. Reutzel, J. Taneja, J. Hsu, and S. Shankar, "An information-centric energy infrastructure: The Berkeley view," *Sustainable Computing: Informatics and Systems*, vol. 1, no. 1, pp. 7–22, Mar. 2011, ISSN: 22105379. DOI: 10.1016/j.suscom.2010.10.001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S2210537910000028.

[34] N. Schuetz, A. Kovaleva, and J. G. Koomey, "eBay Inc.: a case study of the organizational change underlying technical infrastructure optimization," Steyer-Taylor Center for Energy Policy and Finance, Stanford University, Case Study, Sep. 2013. [Online]. Available: http://download1701.mediafire.com/302igezru9sg/8ema554a2ho9ifj/Stanford+eBay+Case+Study-+FINAL-130926.pdf.

[35] C. Belady, "Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric," The Green Grid, White Paper, 2010. [Online]. Available: http://www.thegreengrid.org/~/media/WhitePapers/Carbon%20Usage%20Effectiveness%20White%20Paper_v3.pdf?lang=en.

[36] Global Taskforce, "Harmonizing Global Metrics for Data Center Energy Efficiency," White Paper, Oct. 2012. [Online]. Available: http://iet.jrc.ec.europa.eu/energyefficiency/sites/energyefficiency/files/files/documents/ICT_CoC/harmonizing_global_metrics_for_data_center_energy_efficiency_2012-10-02.pdf.

[37] B. Raghavan and J. Ma, "The energy and emergy of the internet," in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, 2011, p. 9. [Online]. Available: http://dl.acm.org/citation.cfm?id=2070571.

[38] M. Jamalzadeh and N. Behravan, "An exhaustive framework for better data centers' energy efficiency and greenness by using metrics," *Indian Journal of Computer Science and Engineering*, vol. 2, no. 6, pp. 813–822, 2011. [Online]. Available: http://www.doaj.org/doaj?func=fulltext&aId=920933.

[39] V. Avelar, D. Azevedo, and A. French, "PUE: a comprehensive examination of the metric," The Green Grid, White Paper WP-49, 2012. [Online]. Available: http://www.thegreengrid.org/~/media/WhitePapers/WP49-PUE%20A%20Comprehensive%20Examination%20of%20the%20Metric_v6.pdf?lang=en.

[40] R. Miller, *EPA to use PUE in Data Center Energy Star*, Apr. 2009. [Online]. Available: http://www.datacenterknowledge.com/archives/2009/04/22/epa-to-use-pue-in-data-center-energy-star/.

[41] Google, *Efficiency: How we do it*, 2015. [Online]. Available: https://www.google.com/about/datacenters/efficiency/internal/.

[42] Facebook, *Prineville Data Center - PUE/WUE*, Apr. 2016. [Online]. Available: https://www.facebook.com/PrinevilleDataCenter/app/399244020173259/.

[43] Facebook, *Forest City Data Center - PUE/WUE*, Apr. 2016. [Online]. Available: https://www.facebook.com/ForestCityDataCenter/app/288655784601722/.

[44] P. Delforge and J. Whitney, "Data Center Efficiency Assessment," Natural Resources Defense Council, Issue Paper IP:14-08-A, Aug. 2014.

[45] A. Patrizio, *Data Center Managers Worn Out by PUE Chase*, May 2013. [Online]. Available: http://slashdot.org/topic/datacenter/data-center-managers-worn-out-by-pue-chase/.

[46] R. Miller, *Uptime Institute: The Average PUE is 1.8*, May 2011. [Online]. Available: http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/.

[47] Digital Realty Trust, *North America Campos Survey Results*, Jan. 2013. [Online]. Available: https://c.na6.content.force.com/sfc/dist/version/download?oid=00D300000005uRq&ids=06880000000k573&d=/a/80000000CpC7/k_RJOcsv31zvPC4hgEz9NMQjNd0m4KjS_CzGO5_ni48%3D.

[48] J. Beltran, *2014 Data Center Industry Survey*, Nov. 2014. [Online]. Available: https://journal.uptimeinstitute.com/2014-data-center-industry-survey/.

[49] M. Stansberry, "Data Center Industry Survey 2015," Uptime Institute, Tech. Rep., 2015. [Online]. Available: https://uptimeinstitute.com/uptime_assets/08200c5b92224d561ba5ff84523e5fdefeec6b58cbf64c19da7338e185a9c828-survey15.pdf.

[50] T. Roberts, *Are You Suffering from PUE Envy?* May 2013. [Online]. Available: http://www.datacenterknowledge.com/archives/2013/05/29/are-you-suffering-from-pue-envy/.

[51] Greenhouse Gas Protocol, “Guide for assessing GHG emissions of data centers,” in *GHG Protocol Product Life Cycle Accounting and Reporting Standard ICT Sector Guidance*, Mar. 2012. [Online]. Available: http://www.ghgprotocol.org/files/ghgp/Chapter_8_GHGP-ICT%20Data%20Center%20guide%20v2-3%2010MAR2012.pdf.

[52] P. Dijkhuis, *Government regulation for datacenters in Amsterdam*, Mar. 2008. [Online]. Available: http://blog.leaseweb.com/2008/03/30/government-regulation-for-datacenters-in-amsterdam/.

[53] N. Henderson, *EvoSwitch Promotes Amsterdam as European Data Center Gateway*, Aug. 2011. [Online]. Available: http://www.thewhir.com/web-hosting-news/evoswitch-promotes-amsterdam-as-european-data-center-gateway.

[54] D. Cole, “Data center energy efficiency—looking beyond PUE,” No Limits Software, White Paper 4, 2011. [Online]. Available: http://www.nolimitssoftware.com/docs/DataCenterEnergyEfficiency_LookingBeyond.pdf.

[55] S. Shankland, *Google uncloaks once-secret server*, Apr. 2009. [Online]. Available: http://news.cnet.com/8301-1001_3-10209580-92.html.

[56] J. S. Klaus, *Data Center Energy: Past, Present and Future (Part One)*, Article, Sep. 2012. [Online]. Available: http://www.datacenterjournal.com/it/data-center-energy-past-present-and-future-part-one/.

[57] M. Armbrust, O. Fox, R. Griffith, A. D. Joseph, Y. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “Above the clouds: a Berkeley view of cloud computing,” Feb. 2009. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.361.8444.

[58] J. M. Kaplan, W. Forrest, and N. Kindler, “Revolutionizing data center energy efficiency,” McKinsey & Company, Tech. Rep., Jul. 2008. [Online]. Available: http://www.sallan.org/pdf-docs/McKinsey_Data_Center_Efficiency.pdf.

[59] G. Cook, “How Clean is Your Cloud?” Greenpeace International, Tech. Rep., Apr. 2012. [Online]. Available: http://www.greenpeace.org/international/Global/international/publications/climate/2012/iCoal/HowCleanisYourCloud.pdf.

[60] LBNL, *Benchmarking: Data Centers - Case Study Reports*, 2003. [Online]. Available: http://hightech.lbl.gov/dc-benchmarking-results.html.

[61] S. Greenberg, E. Mills, B. Tschudi, P. Rumsey, and B. Myatt, “Best practices for data centers: lessons learned from benchmarking 22 data centers,” *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings in Asilomar, CA. ACEEE, August*, vol. 3, pp. 76–87, 2006. [Online]. Available: http://energy.lbl.gov/EA/mills/EMills/pubs/pdf/aceee-datacenters.pdf.

[62] M. Sheppy, C. Lobato, O. Van Geet, S. Pless, K. Donovan, and C. Powers, “Reducing data center loads for a large-scale, low-energy office building: NREL’s research support facility,” National Renewable Energy Laboratory (NREL), Golden, CO., Tech. Rep., 2011. [Online]. Available: http://www.osti.gov/scitech/biblio/1031393.

[63] G. Cook, “A Clean Energy Road Map for Apple,” Greenpeace International, Tech. Rep. JN 417 UPDATE, Jul. 2012. [Online]. Available: http://www.greenpeace.org/international/Global/international/publications/climate/2012/iCoal/Apple_Clean_Energy_Road_Map.pdf.

[64] J. Rath and B. Kleyman, “Guide to Modular Data Centers,” Data Center Knowledge, Tech. Rep., Mar. 2013. [Online]. Available: http://www.findwhitepapers.com/force-download.php?id=26174.

[65] P. Flynn, “Quantitative analysis of energy and financial savings for full-year operation of modular data center relative to raised floor environment,” IO, Technical Paper, Sep. 2015. [Online]. Available: https://www.io.com/wp-content/uploads/PUE-Technical-Paper.pdf.

[66] Greenhouse Gas Protocol, *GHG emissions from purchased electricity*, Aug. 2012. [Online]. Available: http://www.ghgprotocol.org/files/ghgp/tools/GHG-emissions-from-purchased-electricity(Version-4_4_Aug-2012).xlsx.

[67] U.S. Environmental Protection Agency, *Product Specifications & Partner Commitments Search*, 2016. [Online]. Available: https://www.energystar.gov/products/spec/.

[68] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf, "Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward," in *Proceedings of the ACM SIGMETRICS/International conference on measurement and modeling of computer systems*, Pittsburgh, PA, 2013, pp. 153–166. [Online]. Available: http://dl.acm.org/citation.cfm?id=2465760.

[69] M. Stansberry and J. Kudritzki, *2012 Data Center Industry Survey*, 2012.

[70] E. Masanet, A. Shehabi, L. Ramakrishnan, J. Liang, X. Ma, B. Walker, and V. Hendrix, "The Energy Efficiency Potential of Cloud-Based Software: A U.S. Case Study," Lawrence Berkeley National Laboratory, Berkeley, CA, Tech. Rep., Jun. 2013. [Online]. Available: http://www.olomedia.it/files/pages/cloud_efficiency_study.pdf.

[71] NRDC, "The Carbon Emissions of Server Computing for Small- to Medium-Sized Organizations: A Performance Study of On-Premise vs. The Cloud," Natural Resources Defense Council, Tech. Rep., Oct. 2012. [Online]. Available: http://www.wspenvironmental.com/media/docs/ourlocations/usa/NRDC-WSP_Cloud_Computing.pdf.

[72] J. G. Koomey and P. Flynn, "How to run data center operations like a well oiled machine," *Datacenter Dynamics Focus*, vol. 3, no. 37, p. 81, Oct. 2014. [Online]. Available: http://content.yudu.com/Library/A31vdg/FocusVolume3Issue37/resources/index.htm.

[73] J. G. Koomey, *Three Pillars of Modern Data Center Operations*, Feb. 2016. [Online]. Available: http://www.datacenterknowledge.com/archives/2016/02/02/three-pillars-modern-data-center-operations/.

[74] ENERGY STAR, *ENERGY STAR Certified Data Centers*, 2016. [Online]. Available: https://www.energystar.gov/index.cfm?fuseaction=labeled_buildings.showDataCenters..

[75] NABERS, *Rating Register*, 2016. [Online]. Available: http://www.nabers.gov.au/public/WebPages/ContentStandard.aspx?module=30&template=3&id=310&side=AssessorTertiary.htm.

[76] B. Riley, *CRC Annual Report Publication*, Jan. 2014. [Online]. Available: http://www.environment-agency.gov.uk/business/topics/pollution/146938.aspx.

[77] JRC, *Data Centres Energy Efficiency*, Mar. 2014. [Online]. Available: http://iet.jrc.ec.europa.eu/energyefficiency/ict-codes-conduct/data-centres-energy-efficiency.

[78] L. M. Hilty and B. Aebischer, Eds., *ICT Innovations for Sustainability*, ser. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2015, vol. 310. [Online]. Available: http://link.springer.com/10.1007/978-3-319-09228-7.

[79] E. R. Masanet and H. S. Matthews, "Environmental Applications of Information and Communication Technology [special issue]," *Journal of Industrial Ecology*, vol. 14, no. 5, pp. 685–862, Oct. 2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/jiec.2010.14.issue-5/issuetoc.

[80] "E-Commerce, the Internet, and the Environment [special issue]," *Journal of Industrial Ecology*, vol. 6, no. 2, D. Rejeski, Ed., pp. 1–161, Apr. 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/jiec.2002.6.issue-2/issuetoc.

[81] S. Roberts, "Measuring the relationship between ICT and the environment," 2009. [Online]. Available: http://www.oecd-ilibrary.org/science-and-technology/measuring-the-relationship-between-ict-and-the-environment_221687775423.

[82] A. Auweter, D. Kranzlmüller, A. Tahamtan, A. M. Tjoa, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, and G. Weikum, Eds., *ICT as key technology against global warming*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7453, ISBN: 978-3-642-32605-9. [Online]. Available: http://link.springer.com/10.1007/978-3-642-32606-6.

[83] L. Erdmann, L. Hilty, J. Goodman, and P. Arnfalk, "The Future Impact of ICTs on Environmental Sustainability," Institute for Prospective Technological Studies, Seville, Spain, Technical Report EUR 21384 EN, Aug. 2004. [Online]. Available: http://library.certh.gr/libfiles/PDF/EU-EKETA-1447-THE-FUTURE-IMPACT-EUR-21384en-AUG-2004-PP64.pdf.

[84] M. S. Jørgensen, M. M. Andersen, A. Hansen, H. Wenzel, T. Thoning, U. J. Pedersen, M. Falch, B. Rasmussen, S. I. Olsen, and O. Willum, “Green Technology Foresight about environmentally friendly products and materials,” *Environmental Protection Agency, Denmark.*, 2006. [Online]. Available: http://www2.mst.dk/udgiv/publications/2006/87-7052-216-2/pdf/87-7052-217-0.pdf.

[85] S. Lohr, “Smart infrastructure brings efficiencies to roads, rail, water and food distribution,” *The New York Times*, Apr. 2009, ISSN: 0362-4331. [Online]. Available: http://www.nytimes.com/2009/04/30/business/energy-environment/30smart.html.

[86] S. Ruth, “Reducing ICT-related carbon emissions: an exemplar for global energy policy?” *IETE technical review*, vol. 28, no. 3, pp. 207–211, 2011. [Online]. Available: http://www.tandfonline.com/doi/abs/10.4103/0256-4602.81229.

[87] J. Romm, A. Rosenfeld, and S. Herrmann, “The internet economy and global warming: A scenario of the impact of e-commerce on energy and the environment,” *Center for Energy and Climate Solutions, December, http://www. cool-companies. org/energy/cecs. cfm*, 1999. [Online]. Available: http://infohouse.p2ric.org/ref/04/03784/0378401.pdf.

[88] J. A. Laitner and K. Ehrhardt-Martinez, “Information and Communication Technologies: The Power of Productivity,” American Council for an Energy-Efficient Economy, Washington, D.C., E081, Feb. 2008. [Online]. Available: http://aceee.org/sites/default/files/publications/researchreports/E081.pdf.

[89] N. Elliott, M. Molina, and D. Trombley, “A defining framework for intelligent efficiency,” American Council for an Energy-Efficient Economy, Washington, DC, Tech. Rep. E125, Jun. 2012. [Online]. Available: http://aceee.org/sites/default/files/publications/researchreports/e125.pdf.

[90] S. Seidel and J. Ye, “Leading by example: using information and communication technologies to achieve Federal sustainability goals,” Center for Climate and Energy Solutions, Tech. Rep., Sep. 2012. [Online]. Available: http://www.c2es.org/docUploads/federal-sustainability-ict.pdf.

[91] Accenture, “SMARTer2030: ICT Solutions for 21st Century Challenges,” Tech. Rep., 2015. [Online]. Available: http://smarter2030.gesi.org/downloads/Full_report2.pdf.

[92] R. Rattle, *Computing Our Way to Paradise?: The Role of Internet and Communication Technologies in Sustainable Consumption and Globalization*. Rowman & Littlefield, Jan. 2010, ISBN: 978-0-7591-1933-8.

[93] F. Berkhout and J. Hertin, “De-materialising and re-materialising: digital technologies and the environment,” *Futures*, vol. 36, no. 8, pp. 903–920, Oct. 2004, ISSN: 00163287. DOI: 10.1016/j.futures.2004.01.003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0016328704000047.

[94] M. Börjesson Rivera, C. Håkansson, Å. Svenfelt, and G. Finnveden, “Including second order effects in environmental assessments of ICT,” *Environmental Modelling & Software*, vol. 56, pp. 105–115, Jun. 2014, ISSN: 13648152. DOI: 10.1016/j.envsoft.2014.02.005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1364815214000565.

[95] L. Erdmann and L. M. Hilty, “Scenario analysis: exploring the macroeconomic impacts of information and communication technologies on greenhouse gas emissions,” *Journal of Industrial Ecology*, vol. 14, no. 5, pp. 826–843, Oct. 2010, ISSN: 10881980. DOI: 10.1111/j.1530-9290.2010.00277.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1530-9290.2010.00277.x.

[96] J. G. Koomey, H. S. Matthews, and E. Williams, “Smart everything: will intelligent systems reduce resource use?” *Annual Review of Environment and Resources*, vol. 38, no. 1, pp. 311–343, Oct. 2013, ISSN: 1543-5938, 1545-2050. DOI: 10.1146/annurev-environ-021512-110549. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev-environ-021512-110549.

[97] L. Yi and H. R. Thomas, “A review of research on the environmental impact of e-business and ICT,” *Environment International*, vol. 33, no. 6, pp. 841–849, Aug. 2007, ISSN: 01604120. DOI: 10.1016/j.envint.2007.03.015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0160412007000736.

[98] K. P. Green and A. Mathur, “Measuring and reducing Americans’ indirect energy use,” *AEI Energy and Environment Outlook*, Dec. 2008. [Online]. Available: http://www.aei.org/wp-content/uploads/2011/10/20081204_EEONo2g.pdf.

[99] J. Malmodin, Å. Moberg, D. Lundén, G. Finnveden, and N. Lövehagen, "Greenhouse gas emissions and operational electricity use in the ICT and entertainment & media sectors," *Journal of Industrial Ecology*, vol. 14, no. 5, pp. 770–790, Oct. 2010, ISSN: 10881980. DOI: 10.1111/j.1530-9290.2010.00278.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1530-9290.2010.00278.x.

[100] W. Van Heddeghem, S. Lambert, B. Lannoo, D. Colle, M. Pickavet, and P. Demeester, "Trends in worldwide ICT electricity consumption from 2007 to 2012," *Computer Communications*, vol. 50, pp. 64–76, Sep. 2014, ISSN: 01403664. DOI: 10.1016/j.comcom.2014.02.008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0140366414000619.

[101] L. Norford, A. Hatcher, J. Harris, J. Roturier, and O. Yu, "Electricity use in information technologies," *Annual Review of Energy*, vol. 15, no. 1, pp. 423–453, 1990. [Online]. Available: http://www.annualreviews.org/doi/pdf/10.1146/annurev.eg.15.110190.002231.

[102] J. G. Koomey, M. Piette, M. Cramer, and J. H. Eto, "Efficiency improvements in US office equipment: expected policy impacts and uncertainties," *Energy Policy*, vol. 24, no. 12, pp. 1101–1110, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421596001012.

[103] K. Kawamoto, J. G. Koomey, B. Nordman, R. E. Brown, M. A. Piette, M. Ting, and A. K. Meier, "Electricity used by office equipment and network equipment in the US," *Energy*, vol. 27, no. 3, pp. 255–269, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360544201000846.

[104] K. W. Roth, F. Goldstein, and J. Kleinman, "Energy Consumption by Office and Telecommunications Equipment in Commercial Buildings, Volume I: Energy Consumption Baseline," Arthur D. Little, Inc., Cambridge, MA, Tech. Rep. ADL 7285-00, Jan. 2002. [Online]. Available: http://www.biblioite.ethz.ch/downloads/Roth_ADL_1.pdf.

[105] B. Nordman and A. K. Meier, *Energy Consumption of Home Information Technology*, Oct. 2003. [Online]. Available: http://energy.lbl.gov/controls/_not-used/homeit/.

[106] K. W. Roth, R. Ponoum, and F. Goldstein, "U.S. Residential Information Technology Energy Consumption in 2005 and 2010," TIAX, LLC, Cambridge, MA, Tech. Rep. D0295, Mar. 2006. [Online]. Available: http://www.biblioite.ethz.ch/downloads/residential_information_technology_energy_consumption_2006.pdf.

[107] W. S. Baer, S. Hassell, and B. A. Vollaard, *Electricity requirements for a digital society*. RAND Corporation, 2002.

[108] J. G. Koomey, "Growth in data center electricity use 2005 to 2010," *Oakland, CA: Analytics Press. August*, vol. 1, p. 2010, 2011. [Online]. Available: http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2011finalversion.pdf.

[109] E. Williams, "Environmental effects of information and communications technologies," *Nature*, vol. 479, no. 7373, pp. 354–358, Nov. 2011, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature10682. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature10682.

[110] B. Allenby and D. Unger, "Information technology impacts on the US energy demand profile," *RAND Corporation*, 2001.

[111] I. M. Azevedo, "Consumer end-use energy efficiency and rebound effects," *Annual Review of Environment and Resources*, vol. 39, no. 1, pp. 393–418, Oct. 2014, ISSN: 1543-5938, 1545-2050. DOI: 10.1146/annurev-environ-021913-153558. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev-environ-021913-153558.

[112] G. Young, "Illuminating the Issues: Digital Signage and Philadelphia's Green Future," SCRUB: Public Voice for Public Space, Philadelphia, PA, Tech. Rep., 2013. [Online]. Available: http://www.scenic.org/storage/documents/Digital_Signage_Final_Dec_14_2010.pdf.

[113] L. M. Hilty, P. Arnfalk, L. Erdmann, J. Goodman, M. Lehmann, and P. A. Wäger, "The relevance of information and communication technologies for environmental sustainability – A prospective simulation study," *Environmental Modelling & Software*, vol. 21, no. 11, pp. 1618–1629, Nov. 2006, ISSN: 13648152. DOI: 10.1016/j.envsoft.2006.05.007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1364815206001204.

[114] K. Gillingham, D. Rapson, and G. Wagner, "The rebound effect and energy efficiency policy," 2015. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2550710.

[115] S. Borenstein, "A microeconomic framework for evaluating energy efficiency rebound and some implications," National Bureau of Economic Research, Tech. Rep., 2013. [Online]. Available: http://www.nber.org/papers/w19044.

[116] M. Hesse, "Shipping news: the implications of electronic commerce for logistics and freight transport," *Resources, Conservation and Recycling*, vol. 36, no. 3, pp. 211–240, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921344902000836.

[117] D. Harrington, "From first mile to last mile: Global industrial & logistics trends," Colliers International, Tech. Rep., Oct. 2015. [Online]. Available: http://www.colliers.com/-/media/files/marketresearch/global/2015-global-reports/global-logistics-2015.pdf.

[118] Shorr Packaging Corp, *The Amazon effect: impacts on shipping and retail*, Jun. 2015. [Online]. Available: http://www.shorr.com/packaging-news/2015-06/amazon-effect-impacts-shipping-and-retail.

[119] A. M. Mohan, "E-commerce packaging pitfalls & opportunities," *Packaging World*, Dec. 2014. [Online]. Available: http://www.packworld.com/trends-and-issues/distribution/e-commerce-packaging-pitfalls-opportunities.

[120] L. A. Greening, D. L. Greene, and C. Difiglio, "Energy efficiency and consumption – the rebound effect – a survey," *Energy Policy*, vol. 28, pp. 389–401, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421500000215.

[121] A. Plepys, "The grey side of ICT," *Environmental Impact Assessment Review*, vol. 22, no. 5, pp. 509–523, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0195925502000252.

[122] J. G. Bull and R. A. Kozak, "Comparative life cycle assessments: the case of paper and digital media," *Environmental Impact Assessment Review*, vol. 45, pp. 10–18, Feb. 2014, ISSN: 01959255. DOI: 10.1016/j.eiar.2013.10.001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0195925513000942.

[123] J. H. Schmidt and M. Pizzol, "Critical review of four comparative lifecycle assessments of printed and electronic communication," Denmark, Tech. Rep., Dec. 2014. [Online]. Available: http://lca-net.com/files/Review_of_four_LCAs_on_printed_versus_electronic_media.pdf.

[124] European Commission Directorate-General for the Information Society and Media, *ICT and energy efficiency: the case for manufacturing. Recommendations of the consultation group*. Luxembourg: EUR-OP, 2009, OCLC: 847296210, ISBN: 978-92-79-11306-2.

[125] S. Obayashi, "Multidisciplinary design optimization of aircraft wing planform based on evolutionary algorithms," in *IEEE International Conference on Systems Man and Cybernetics*, vol. 4, INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE), 1998, pp. 3148–3153. [Online]. Available: https://www.researchgate.net/profile/Shigeru_Obayashi/publication/3776439_Multidisciplinary_design_optimization_of_aircraft_wing_planformbased_on_evolutionary_algorithms/links/00463516babd785fcd000000.pdf.

[126] A. Keane and P. Nair, *Computational Approaches for Aerospace Design: The Pursuit of Excellence*, 1 edition. Chichester, England ; Hoboken, N.J: Wiley, Aug. 2005, ISBN: 978-0-470-85540-9.

[127] J. Sobieszczanski-Sobieski and R. T. Haftka, "Multidisciplinary aerospace design optimization: survey of recent developments," *Structural optimization*, vol. 14, no. 1, pp. 1–23, 1997, ISSN: 1615-1488. DOI: 10.1007/BF01197554. [Online]. Available: http://dx.doi.org/10.1007/BF01197554.

[128] G. J. Kennedy, G. K. W. Kenway, and J. Martins, "High aspect ratio wing design: optimal aerostructural tradeoffs for the next generation of materials," in *Proceedings of the AIAA Science and Technology Forum and Exposition (SciTech), National Harbor, MD*, 2014. [Online]. Available: http://arc.aiaa.org/doi/pdf/10.2514/6.2014-0596.

[129] M. Hobday, A. Davies, and A. Prencipe, "Systems integration: a core capability of the modern corporation," *Industrial and Corporate Change*, vol. 14, no. 6, pp. 1109–1143, Nov. 2005. [Online]. Available: http://icc.oxfordjournals.org/content/14/6/1109.abstract.

[130] H. Siikavirta, M. Punakivi, M. Kärkkäinen, and L. Linnanen, "Effects of e-commerce on greenhouse gas emissions: a case study of grocery home delivery in Finland," *Journal of industrial ecology*, vol. 6, no. 2, pp. 83–97, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1162/108819802763471807/abstract.

[131] Airbus, *Simulation and tests*, 2016. [Online]. Available: http://www.airbus.com/innovation/proven-concepts/in-design/simulation-and-tests/.

[132] J. Basbagill, F. Flager, M. Lepech, and M. Fischer, "Application of life-cycle assessment to early stage building design for reduced embodied environmental impacts," *Building and Environment*, vol. 60, pp. 81–92, Feb. 2013, ISSN: 03601323. DOI: 10.1016/j.buildenv.2012.11.009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0360132312003071.

[133] J. Kim, M. Xu, R. Kahhat, B. Allenby, and E. Williams, "Design and assessment of a sustainable networked system in the US; case study of book delivery system," in *IEEE International Symposium on Electronics and the Environment*, IEEE, 2008, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4562874.

[134] E. Williams and T. Tagami, "Energy use in sales and distribution via e-commerce and conventional retail: a case study of the Japanese book sector," *Journal of Industrial Ecology*, vol. 6, no. 2, pp. 99–114, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1162/108819802763471816/abstract.

[135] H. S. Matthews, E. Williams, T. Tagami, and C. T. Hendrickson, "Energy implications of online book retailing in the United States and Japan," *Environmental Impact Assessment Review*, vol. 22, no. 5, pp. 493–507, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0195925502000240.

[136] H. S. Matthews, C. T. Hendrickson, and D. Soh, "The net effect: environmental implications of e-commerce and logistics," in *Electronics and the Environment, 2001. Proceedings of the 2001 IEEE International Symposium on*, IEEE, 2001, pp. 191–195. [Online]. Available: http://www.cmu.edu/gdi/docs/environmental-and-economic.pdf.

[137] H. S. Matthews, C. T. Hendrickson, and D. L. Soh, "Environmental and economic effects of e-commerce: a case study of book publishing and retail logistics," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1763, no. 1, pp. 6–12, 2001. [Online]. Available: http://trb.metapress.com/index/8535166HQ111J423.pdf.

[138] D. Sivaraman, S. Pacca, K. Mueller, and J. Lin, "Comparative energy, environmental, and economic analysis of traditional and e-commerce DVD rental networks," *Journal of Industrial Ecology*, vol. 11, no. 3, pp. 77–91, 2007. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1162/jiec.2007.1240/abstract.

[139] A. Shehabi, B. Walker, and E. Masanet, "The energy and greenhouse-gas implications of internet video streaming in the United States," *Environmental Research Letters*, vol. 9, no. 5, p. 054 007, May 2014, ISSN: 1748-9326. DOI: 10.1088/1748-9326/9/5/054007. [Online]. Available: http://stacks.iop.org/1748-9326/9/i=5/a=054007?key=crossref.109059040b415b8f966a41991d5465e4.

[140] C. Weber, J. G. Koomey, and H. S. Matthews, "The energy and climate change implications of different music delivery methods," *Journal of Industrial Ecology*, vol. 14, no. 5, pp. 754–769, Oct. 2010, ISSN: 10881980. DOI: 10.1111/j.1530-9290.2010.00269.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1530-9290.2010.00269.x.

[141] C. Weber, C. Hendrickson, P. Jaramillo, S. Matthews, A. Nagengast, and R. Nealer, "Life cycle comparison of traditional retail and e-commerce logistics for electronic products: a case study of buy.com," *Green Design Institute, Carnegie Mellon University*, 2008. [Online]. Available: http://www.cmu.edu/gdi/docs/life-cycle-comparison.pdf.

[142] J. B. Edwards, A. C. McKinnon, and S. L. Cullinane, "Comparative analysis of the carbon footprints of conventional and online retailing: A "last mile" perspective," *International Journal of Physical Distribution & Logistics Management*, vol. 40, no. 1/2, Á. Halldórsson, Ed., pp. 103–123, Feb. 2010, ISSN: 0960-0035. DOI: 10.1108/09600031011018055. [Online]. Available: http://www.emeraldinsight.com/doi/abs/10.1108/09600031011018055.

[143] H. S. Matthews and C. T. Hendrickson, "The economic and environmental implications of centralized stock keeping," *Journal of Industrial Ecology*, vol. 6, no. 2, pp. 71–81, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1162/108819802763471799/abstract.

[144] D. L. Gard and G. A. Keoleian, "Digital versus print: energy performance in the selection and use of scholarly journals," *Journal of Industrial Ecology*, vol. 6, no. 2, pp. 115–132, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1162/108819802763471825/abstract.

[145] A. Seetharam, M. Somasundaram, D. Towsley, J. Kurose, and P. Shenoy, "Shipping to streaming: is this shift green?" In *Proceedings of the first ACM SIGCOMM workshop on Green networking*, ACM, 2010, pp. 61–68. [Online]. Available: http://dl.acm.org/citation.cfm?id=1851304.

[146] K. Mayers, J. Koomey, R. Hall, M. Bauer, C. France, and A. Webb, "The carbon footprint of games distribution," *Journal of Industrial Ecology*, n/a–n/a, Aug. 2014, ISSN: 10881980. DOI: 10.1111/jiec.12181. [Online]. Available: http://doi.wiley.com/10.1111/jiec.12181.

[147] Å. Moberg, M. Johansson, G. Finnveden, and A. Jonsson, "Printed and tablet e-paper newspaper from an environmental perspective — a screening life cycle assessment," *Environmental Impact Assessment Review*, vol. 30, no. 3, pp. 177–191, Apr. 2010, ISSN: 01959255. DOI: 10.1016/j.eiar.2009.07.001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0195925509000936.

[148] M. Zurkirch and I. Reichart, "Environmental impacts of telecommunication services," *Greener Management International*, no. 32, pp. 70–88, 2000. [Online]. Available: http://www.ingentaconnect.com/content/glbj/ene/2002/00000001/00000108/art00012.

[149] B. Aebischer and A. Huser, "Networking in private households: impacts on electricity consumption," Swiss Federal Office of Energy, Tech. Rep., 2000. [Online]. Available: http://www.bfe.admin.ch/php/modules/enet/streamfile.php?file=000000006772_01.pdf.

[150] R. Atkyns, M. Blazek, and J. Roitz, "Measurement of environmental impacts of telework adoption amidst change in complex organizations: AT&T survey methodology and results," *Resources, conservation and recycling*, vol. 36, no. 3, pp. 267–285, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921344902000824.

[151] P. Hopkinson and P. James, "UK report on national SUSTEL fieldwork," *Sustainable Telework*, 2003. [Online]. Available: http://webfarm.userve.net/~flexiworker/pdf/Case%20studies.pdf.

[152] H. S. Matthews and E. Williams, "Telework adoption and energy use in building and transport sectors in the United States and Japan," *Journal of infrastructure systems*, vol. 11, no. 1, pp. 21–30, Mar. 2005. [Online]. Available: http://www.cmu.edu/gdi/docs/telework-adoption.pdf.

[153] K. W. Roth, T. Rhodes, and R. Ponoum, "The energy and greenhouse gas emission impacts of telecommuting in the US," in *IEEE International Symposium on Electronics and the Environment*, IEEE, 2008, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4562945.

[154] E. Kitou and A. Horvath, "External air pollution costs of telework," *The International Journal of Life Cycle Assessment*, vol. 13, no. 2, pp. 155–165, Mar. 2008, ISSN: 0948-3349, 1614-7502. DOI: 10.1065/lca2007.06.338. [Online]. Available: http://www.springerlink.com/index/10.1065/lca2007.06.338.

[155] A. Brown, J. Gonder, and B. Repac, "An analysis of possible energy impacts of automated vehicle," in *Road Vehicle Automation*, G. Meyer and S. Beiker, Eds., Cham: Springer International Publishing, 2014, pp. 137–153, ISBN: 978-3-319-05989-1. [Online]. Available: http://link.springer.com/10.1007/978-3-319-05990-7_13.

[156] T. Langer and S. Vaidyanathan, "Smart Freight: Applications of Information and Communications Technologies to Freight System Efficiency," Washington, DC: American Council for an Energy-Efficient Economy, White Paper, 2014. [Online]. Available: http://www.indiaenvironmentportal.org.in/files/file/Smart%20Freight.pdf.

[157] R. J. Meyers, E. D. Williams, and H. S. Matthews, "Scoping the potential of monitoring and control technologies to reduce energy use in homes," *Energy and Buildings*, vol. 42, no. 5, pp. 563–569, May 2010, ISSN: 03787788. DOI: 10.1016/j.enbuild.2009.10.026. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0378778809002758.

[158] F. Mattern, T. Staake, and M. Weiss, "ICT for green: how computers can help us to conserve energy," in *Proceedings of the 1st international conference on energy-efficient computing and networking*, ACM, 2010, pp. 1–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=1791316.

[159] A. L. Davis, T. Krishnamurti, B. Fischhoff, and W. Bruine de Bruin, "Setting a standard for electricity pilot studies," *Energy Policy*, vol. 62, pp. 401–409, Nov. 2013, ISSN: 03014215. DOI: 10.1016/j.enpol.2013.07.093. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0301421513007362.

[160] D. Schwartz, B. Fischhoff, T. Krishnamurti, and F. Sowell, "The Hawthorne effect and energy awareness," *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. 15 242–15 246, Sep. 2013, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1301687110. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1301687110.

[161] E. A. Rogers, R. N. Elliott, S. Kwatra, D. Trombley, and V. Nadadur, "Intelligent Efficiency: Opportunities, Barriers, and Solutions," American Council for an Energy-Efficient Economy, Washington, D.C., Tech. Rep. E13J, Oct. 2013.

[162] U.S. Department of Energy, "Advanced sensors, control, platforms, and modeling for manufacturing (smart manufacturing): Technology assessment," U.S. Dept. of Energy, Quadrennial Technology Review, Feb. 2015. [Online]. Available: http://energy.gov/sites/prod/files/2015/02/f19/QTR%20Ch8%20-%20Smart%20Manufacturing%20TA%20Feb-13-2015.pdf.

[163] E. Masanet, "Energy benefits of electronic controls at small and medium sized U.S. manufacturers," *Journal of Industrial Ecology*, vol. 14, no. 5, pp. 696–702, Oct. 2010, ISSN: 10881980. DOI: 10.1111/j.1530-9290.2010.00286.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1530-9290.2010.00286.x.

[164] K. Bunse, M. Vodicka, P. Schönsleben, M. Brülhart, and F. O. Ernst, "Integrating energy efficiency performance in production management – gap analysis between industrial needs and scientific literature," *Journal of Cleaner Production*, vol. 19, no. 6-7, pp. 667–679, Apr. 2011, ISSN: 09596526. DOI: 10.1016/j.jclepro.2010.11.011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0959652610004452.

[165] J. R. Duflou, J. W. Sutherland, D. Dornfeld, C. Herrmann, J. Jeswiet, S. Kara, M. Hauschild, and K. Kellens, "Towards energy and resource efficient manufacturing: A processes and systems approach," *CIRP Annals-Manufacturing Technology*, vol. 61, no. 2, pp. 587–609, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0007850612002016.

[166] J. Davis, T. Edgar, J. Porter, J. Bernaden, and M. Sarli, "Smart manufacturing, manufacturing intelligence and demand-dynamic performance," *Computers & Chemical Engineering*, vol. 47, pp. 145–156, Dec. 2012, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2012.06.037. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0098135412002219.

[167] Smart Manufacturing Leadership Coalition, "Implementing 21st Century Smart Manufacturing," Smart Manufacturing Leadership Coalition, Workshop Summary Report, Jun. 2011. [Online]. Available: https://smartmanufacturingcoalition.org/sites/default/files/implementing_21st_century_smart_manufacturing_report_2011_0.pdf.

[168] P. Palensky and D. Dietrich, "Demand side management: demand response, intelligent energy systems, and smart loads," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, Aug. 2011, ISSN: 1551-3203, 1941-0050. DOI: 10.1109/TII.2011.2158841. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5930335.

[169] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher, "GreenGPS: a participatory sensing fuel-efficient maps application," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, ACM, 2010, pp. 151–164. [Online]. Available: http://dl.acm.org/citation.cfm?id=1814450.

[170] E. Ericsson, H. Larsson, and K. Brundell-Freij, "Optimizing route choice for lowest fuel consumption – Potential effects of a new driver support tool," *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 6, pp. 369–383, Dec. 2006, ISSN: 0968090X. DOI: 10.1016/j.trc.2006.10.001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0968090X06000799.

[171] J. D. Gonder, "Route-based control of hybrid electric vehicles," SAE Technical Paper, Tech. Rep., 2008. [Online]. Available: http://papers.sae.org/2008-01-1315/.

[172] W. Huang, D. M. Bevly, S. Schnick, and X. Li, “Using 3d road geometry to optimize heavy truck fuel efficiency,” in *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, IEEE, 2008, pp. 334–339. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4732656.

[173] H. Doukas, K. D. Patlitzianas, K. Iatropoulos, and J. Psarras, “Intelligent building energy management system using rule sets,” *Building and Environment*, vol. 42, no. 10, pp. 3562–3569, Oct. 2007, ISSN: 03601323. DOI: 10.1016/j.buildenv.2006.10.024. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S036013230600312X.

[174] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, “Occupancy-driven energy management for smart building automation,” in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, ACM, 2010, pp. 1–6. [Online]. Available: http://dl.acm.org/citation.cfm?id=1878433.

[175] Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng, “Duty-cycling buildings aggressively: The next frontier in HVAC control,” in *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, IEEE, 2011, pp. 246–257. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5779042.

[176] P. Henderson and M. Waitner, “Real-time energy management: A case study of three large commercial buildings in Washington, D.C.,” Natural Resources Defense Council, Tech. Rep., Oct. 2013. [Online]. Available: http://www.nrdc.org/business/casestudies/files/tower-companies-case-study.pdf.

[177] D. Z. Sui and D. W. Rejeski, “Environmental impacts of the emerging digital economy: the e-for-environment e-commerce?” *Environmental Management*, vol. 29, no. 2, pp. 155–163, Feb. 2002, ISSN: 0364-152X, 1432-1009. DOI: 10.1007/s00267-001-0027-X. [Online]. Available: http://link.springer.com/10.1007/s00267-001-0027-X.

[178] P. L. Mokhtarian, “If telecommunication is such a good substitute for travel, why does congestion continue to get worse?” *Transportation Letters: The International Journal of Transportation Research*, vol. 1, no. 1, pp. 1–17, Jan. 2009, ISSN: 1942-7867, 1942-7875. DOI: 10.3328/TL.2009.01.01.1-17. [Online]. Available: http://www.tandfonline.com/doi/abs/10.3328/TL.2009.01.01.1-17.

[179] S. Sorrell, “Jevons’ Paradox revisited: the evidence for backfire from improved energy efficiency,” *Energy Policy*, vol. 37, no. 4, pp. 1456–1469, Apr. 2009, ISSN: 03014215. DOI: 10.1016/j.enpol.2008.12.003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0301421508007428.

[180] C. Gossart, “Rebound effects and ICT: a review of the literature,” in *ICT Innovations for Sustainability*, ser. Advances in Intelligent Systems and Computing, L. M. Hilty and B. Aebischer, Eds., vol. 310, Springer International Publishing, 2015, pp. 435–448, ISBN: 978-3-319-09227-0. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-09228-7_26.

[181] P. L. Mokhtarian, “A synthetic approach to estimating the impacts of telecommuting on travel,” *Urban Studies*, vol. 35, no. 2, pp. 215–241, Feb. 1998. [Online]. Available: http://usj.sagepub.com/content/35/2/215.abstract.

[182] S. Choo, P. L. Mokhtarian, and I. Salomon, “Does telecommuting reduce vehicle-miles traveled? An aggregate time series analysis for the US,” *Transportation*, vol. 32, no. 1, pp. 37–64, 2005. [Online]. Available: http://link.springer.com/article/10.1007/s11116-004-3046-7.

[183] J. A. Laitner, “Information technology and US energy consumption: energy hog, productivity tool, or both?” *Journal of Industrial Ecology*, vol. 6, no. 2, pp. 13–24, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1162/108819802763471753/abstract.

[184] J. A. Laitner, J. G. Koomey, E. Worrell, and E. Gumerman, “Re-estimating the annual energy outlook 2000 forecast using updated assumptions about the information economy,” Tech. Rep. LBNL-46418, Jan. 2001. [Online]. Available: http://enduse.lbl.gov/Info/LBNL-46418.pdf.

[185] H. Ishida, “The effect of ICT development on economic growth and energy consumption in Japan,” *Telematics and Informatics*, vol. 32, no. 1, pp. 79–88, Feb. 2015, ISSN: 07365853. DOI: 10.1016/j.tele.2014.04.003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0736585314000331.

[186] K. Takase and Y. Murota, "The impact of IT investment on energy: Japan and US comparison in 2010," *Energy Policy*, vol. 32, no. 11, pp. 1291–1301, Jul. 2004, ISSN: 03014215. DOI: 10.1016/S0301-4215(03)00097-1. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0301421503000971.

[187] S. Murtishaw and L. Schipper, "Disaggregated analysis of US energy consumption in the 1990s: evidence of the effects of the internet and rapid economic growth," *Energy Policy*, vol. 29, no. 15, pp. 1335–1356, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421501000933.

[188] Y. Cho, J. Lee, and T.-Y. Kim, "The impact of ICT investment and energy price on industrial electricity demand: dynamic growth model approach," *Energy Policy*, vol. 35, no. 9, pp. 4730–4738, Sep. 2007, ISSN: 03014215. DOI: 10.1016/j.enpol.2007.03.030. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0301421507001395.

[189] D. Rejeski, "Electronic impact," *The Environmental Forum*, vol. 16, no. 4, pp. 32–38, Aug. 1999.

[190] R. D. Atkinson and A. S. McKay, "Digital Prosperity: Understanding the Economic Benefits of the Information Technology Revolution," ITIF, Washington, D.C., Tech. Rep., Mar. 2007. [Online]. Available: http://www.itif.org/files/digital_prosperity.pdf.

[191] J. G. Koomey, *5 ways to harness info tech to fight climate change*, Feb. 2012. [Online]. Available: https://gigaom.com/2012/02/16/5-ways-to-harness-information-technology-to-fight-climate-change/.

[192] L. M. Hilty, V. C. Coroama, M. O. de Eicker, T. F. Ruddy, and E. Müller, "The Role of ICT in Energy Consumption and Energy Efficiency," EMPA, St. Gallen, Switzerland, Tech. Rep. FP7-ICT-2007-2, 2009. [Online]. Available: http://library.eawag-empa.ch/empa_publications_2009_open_access/EMPA20090243.pdf.

[193] Lawrence Berkeley National Laboratory, *Center of Expertise for Energy Efficiency in Data Centers*, 2016. [Online]. Available: https://datacenters.lbl.gov/.

[194] U.S. Environmental Protection Agency, *Criteria Air Pollutants*, Policies and Guidance, Mar. 2016. [Online]. Available: https://www.epa.gov/criteria-air-pollutants.

[195] D. Aikema, R. Simmonds, and H. Zareipour, "Data centres in the ancillary services market," IEEE, Jun. 2012, pp. 1–10, ISBN: 978-1-4673-2154-9. DOI: 10.1109/IGCC.2012.6322252. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6322252.

[196] Z. Liu, A. Wierman, Y. Chen, B. Razon, and N. Chen, "Data center demand response: Avoiding the coincident peak via workload shifting and local generation," *Performance Evaluation*, vol. 70, no. 10, pp. 770–791, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166531613000928.

[197] B. Aksanli and T. Rosing, "Providing regulation services and managing data center peak power budgets," in *Proceedings of the conference on Design, Automation & Test in Europe*, European Design and Automation Association, 2014, p. 143. [Online]. Available: http://dl.acm.org/citation.cfm?id=2616782.

[198] G. Ghatikar, V. Ganti, N. Matson, and M. A. Piette, "Demand Response Opportunities and Enabling Technologies for Data Centers: Findings from Field Studies," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-5763E, Aug. 2012. [Online]. Available: http://drrc.lbl.gov/sites/all/files/LBNL-5763E.pdf.

[199] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Q. Dang, and K. Pentikousis, "Energy-efficient cloud computing," *The Computer Journal*, vol. 53, no. 7, pp. 1045–1051, Sep. 2010, ISSN: 0010-4620, 1460-2067. DOI: 10.1093/comjnl/bxp080. [Online]. Available: http://comjnl.oxfordjournals.org/cgi/doi/10.1093/comjnl/bxp080.

[200] A. Rahman, X. Liu, and F. Kong, "A survey on geographic load balancing based data center power management in the smart grid environment," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 214–233, 2014, ISSN: 1553-877X. DOI: 10.1109/SURV.2013.070813.00183. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6578864.

[201] F. Kong and X. Liu, "A survey on green-energy-aware power management for datacenters," *ACM Computing Surveys*, vol. 47, no. 2, pp. 1–38, Nov. 2014, ISSN: 03600300. DOI: 10.1145/2642708. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2658850.2642708.

[202] J. Doyle, D. O'Mahony, and R. Shorten, "Server selection for carbon emission control," in *Proceedings of the 2nd ACM SIGCOMM workshop on Green networking*, ACM, 2011, pp. 1–6. [Online]. Available: http://dl.acm.org/citation.cfm?id=2018538.

[203] J. He, X. Deng, D. Wu, Y. Wen, and D. Wu, “Socially-responsible load scheduling algorithms for sustainable data centers over smart grid,” in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, IEEE, 2012, pp. 406–411. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6486018.

[204] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, “Cutting the electric bill for internet-scale systems,” *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, p. 123, Aug. 2009, ISSN: 01464833. DOI: 10.1145/1594977.1592584. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1594977.1592584.

[205] K. Le, R. Bianchini, M. Martonosi, and T. D. Nguyen, “Cost-and energy-aware load distribution across data centers,” *Proceedings of HotPower*, pp. 1–5, 2009. [Online]. Available: http://seelab.ucsd.edu/virtualefficiency/related_papers/27_hotpower09.pdf.

[206] Y. Zhang, Y. Wang, and X. Wang, “Greenware: Greening cloud-scale data centers to maximize the use of renewable energy,” in *ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing*, Springer, 2011, pp. 143–164. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-25821-3_8.

[207] Z. Liu, “Greening geographical load balancing,” Master of Science, California Institute of Technology, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1993767.

[208] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, “It’s not easy being green,” *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 211–222, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2377719.

[209] S. P. Holland and E. T. Mansur, “Is real-time pricing green? The environmental impacts of electricity demand variance,” *The Review of Economics and Statistics*, vol. 90, no. 3, pp. 550–561, 2008. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/rest.90.3.550.

[210] Akamai, *Content Distribution Network*, 2016. [Online]. Available: https://www.akamai.com/us/en/resources/content-distribution-network.jsp.

[211] Cisco, “Cisco Visual Networking Index: Forecast and Methodology, 2015-2020,” White Paper, 2016. [Online]. Available: http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf.

[212] Sandvine, “Global Internet phenomena: Africa, Middle East, & North America,” Sandvine, Waterloo, Ontario, Tech. Rep., Dec. 2015. [Online]. Available: https://www.sandvine.com/downloads/general/global-internet-phenomena/2015/global-internet-phenomena-africa-middle-east-and-north-america.pdf.

[213] Cisco, “Cisco Visual Networking Index: Forecast and Methodology, 2014-2019,” White Paper, May 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf.

[214] E. Limer, *This Box Can Hold an Entire Netflix*, Jul. 2014. [Online]. Available: http://gizmodo.com/this-box-can-hold-an-entire-netflix-1592590450.

[215] Netflix, *Netflix Open Connect*, 2016. [Online]. Available: https://openconnect.netflix.com/en/.

[216] F. Qiu, S. Ahmed, S. S. Dey, and L. Wolsey, “Covering linear programs with violations,” *INFORMS Journal on Computing*, vol. 26, pp. 531–546, 2014. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/ijoc.2013.0582.

[217] J. Malmodin, D. Lundén, Å. Moberg, G. Andersson, and M. Nilsson, “Life cycle assessment of ICT: carbon footprint and operational electricity use from the operator, national, and subscriber perspective in Sweden,” *Journal of Industrial Ecology*, vol. 18, no. 6, pp. 829–845, Dec. 2014, ISSN: 10881980. DOI: 10.1111/jiec.12145. [Online]. Available: http://doi.wiley.com/10.1111/jiec.12145.

[218] D. Costenaro and A. Duer, “The megawatts behind your megabytes: going from data-center to desktop,” *Proceedings of the 2012 ACEEE Summer Study on Energy Efficiency in Buildings, ACEEE, Washington*, pp. 13–65, 2012.

[219] N. Hunt, *Netflix Streaming - More Energy Efficient than Breathing*, May 2015. [Online]. Available: http://techblog.netflix.com/2015/05/netflix-streaming-more-energy-efficient.html.

[220] C. Cain and J. Lesser, “A common sense guide to wholesale electric markets,” Bates White, Tech. Rep., Apr. 2007. [Online]. Available: http://www.bateswhite.com/media/publication/55_media.741.pdf.

[221] Federal Energy Regulatory Commission, *Form 714 - Annual Electric Balancing Authority Area and Planning Area Report*, Aug. 2015. [Online]. Available: http://www.ferc.gov/docs-filing/forms/form-714/data.asp.

[222] MISO, *MISO Second Tier Interface Commercial Pricing Node Definition Changes*, Mar. 2016. [Online]. Available: https://www.misoenergy.org/Library/Repository/Meeting%20Material/Stakeholder/MSC/2016/20160301/20160301%20MSC%20Item%2004g%20Second%20Tier%20Interface%20Cpnode%20Definition%20Changes.pdf.

[223] N. Tacka, *Update: RTO/ISO Regulation Market Comparison*, Feb. 2016. [Online]. Available: http://www.pjm.com/~/media/committees-groups/task-forces/rmistf/20160323/20160323-item-03-update-rto-iso-benchmarking.ashx.

[224] ISO New England, *Pricing Reports*, May 2016. [Online]. Available: http://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info.

[225] New York ISO, *Pricing Data*, May 2016. [Online]. Available: http://www.nyiso.com/public/markets_operations/market_data/pricing_data/index.jsp.

[226] PJM, *Locational Marginal Pricing*, May 2016. [Online]. Available: https://dataminer.pjm.com/dataminerui/pages/public/lmp.jsf.

[227] Midcontinent Independent System Operator, *Archived Real-Time Final Market LMPs*, May 2016. [Online]. Available: https://www.misoenergy.org/Library/MarketReports/Pages/ArchivedRealTimeFinalMarketLMPs.aspx.

[228] Southwest Power Pool, *LMP By Location*, May 2016. [Online]. Available: https://marketplace.spp.org/web/guest/lmp-by-location1.

[229] Electric Reliability Council of Texas, *Historical RTM Load Zone and Hub Prices*, May 2016. [Online]. Available: http://mis.ercot.com/misapp/GetReports.do?reportTypeId=13061&reportTitle=Historical%20RTM%20Load%20Zone%20and%20Hub%20Prices&showHTMLView=&mimicKey.

[230] California ISO, *Open Access Same-time Information System (OASIS)*, May 2016. [Online]. Available: http://oasis.caiso.com/.

[231] PJM, “A review of generation compensation and cost elements in the PJM markets,” Tech. Rep., 2009. [Online]. Available: http://www.pjm.com/~/media/committees-groups/committees/mrc/20100120/20100120-item-02-review-of-generation-costs-and-compensation.ashx.

[232] P. M. Sotkiewicz, *PJM Markets Report*, Jan. 2015. [Online]. Available: https://www.pjm.com/~/media/committees-groups/committees/mc/20150120-webinar/20150120-item-09a-markets-report.ashx.

[233] ISO New England, *Average Monthly Wholesale Load Cost*, May 2016. [Online]. Available: http://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/monthly-wholesale-load-cost-report.

[234] D. B. Patton, P. Van Schaick, and J. Chen, “2014 State of the Market Report for the New York ISO Markets,” Potomac Economics, Tech. Rep., May 2015. [Online]. Available: http://www.nyiso.com/public/webdocs/markets_operations/documents/Studies_and_Reports/Reports/Market_Monitoring_Unit_Reports/2014/NYISO2014SOMReport__5-13-2015_Final.pdf.

[235] D. B. Patton, P. Van Schaick, and J. Chen, “2011 State of the Market Report for the New York ISO Markets,” Potomac Economics, Tech. Rep., Apr. 2012. [Online]. Available: https://www.potomaceconomics.com/uploads/nyiso_reports/NYISO_2011_SOM_Report-Final_4-18-12.pdf.

[236] CAISO, “2015 Annual Report on Market Issues and Performance,” Tech. Rep., May 2016. [Online]. Available: http://caiso.com/Documents/2015AnnualReportonMarketIssuesandPerformance.pdf.

[237] SPP Market Monitoring Unit, “2014 state of the market,” Tech. Rep., Aug. 2015. [Online]. Available: https://www.spp.org/documents/29399/2014%20state%20of%20the%20market%20report.pdf.

[238] ICF International, “Documentation of the Retail Price Model,” Tech. Rep., Jun. 2014. [Online]. Available: https://www.epa.gov/sites/production/files/2015-08/documents/documentation_of_the_retail_price_model.pdf.

[239] U.S. Energy Information Administration, “Annual Energy Outlook 2015,” Tech. Rep. DOE/EIA-0383(2015), Apr. 2015. [Online]. Available: https://www.eia.gov/forecasts/aeo/.

[240] U.S. Energy Information Administration, *Detailed preliminary EIA-826 monthly survey data (back to 1990)*, 2016. [Online]. Available: https://www.eia.gov/electricity/data.cfm.

[241] R. Thomas, *How to Really Reduce New York's Electricity Costs New York AREA*, May 2014. [Online]. Available: http://area-alliance.org/index.php/resources/issue-briefs/how-to-really-reduce-new-yorks-electricity-costs-politicians-should-aim-high-and-focus-on-the-real-cost-drivers-of-new-york-electric-costs-and-its-not-the-lower-hudson-valley-cap/.

[242] L. B. Lave and J. Apt, “Electricity Options for Large Industrial Customers in Western PA,” Carnegie Mellon Electricity Industry Center, Pittsburgh, PA, Report to the Allegheny Conference on Community Development, Aug. 2005. [Online]. Available: https://wpweb2.tepper.cmu.edu/apt/papers/Reports/Electricity%20Options%20for%20Large%20Industrial%20Customers%20in%20Western%20PA%202005.pdf.

[243] B. Sanderson, “For some users, cheap electricity in high-priced New York,” *Politico*, May 2015. [Online]. Available: http://www.politico.com/states/new-york/albany/story/2015/05/for-some-users-cheap-electricity-in-high-priced-new-york-088975.

[244] G. Barbose, C. Goldman, and B. Neenan, “A Survey of Utility Experience with Real Time Pricing,” Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-54238, Dec. 2004. [Online]. Available: https://emp.lbl.gov/sites/all/files/REPORT%20lbnl%20-%2054238.pdf.

[245] ComEd, *Live Prices*, May 2016. [Online]. Available: https://hourlypricing.comed.com/live-prices/.

[246] S. Braithwait, D. Hansen, and M. O'Sheasy, “Retail Electricity Pricing and Rate Design in Evolving Markets,” Edison Electric Institute, Tech. Rep., Jul. 2007. [Online]. Available: http://eei.org/issuesandpolicy/stateregulation/Documents/Retail_Electricity_Pricing.pdf.

[247] S. Borenstein, “Electricity pricing that reflects its real-time cost,” National Bureau of Economic Research, Research Summary, 2009. [Online]. Available: http://www.nber.org/reporter/2009number1/borenstein.html.

[248] U.S. Energy Information Administration, *Negative wholesale electricity prices occur in RTOs*, Jun. 2012. [Online]. Available: http://www.eia.gov/todayinenergy/detail.cfm?id=6730.

[249] U.S. Environmental Protection Agency, *Air Markets Program Data*, 2016. [Online]. Available: https://ampd.epa.gov/ampd/.

[250] U.S. Environmental Protection Agency, *2011 National Emissions Inventory (NEI) Data*, Policies and Guidance, 2011. [Online]. Available: https://www.epa.gov/air-emissions-inventories/2011-national-emissions-inventory-nei-data.

[251] U.S. Environmental Protection Agency, *Emissions & Generation Resource Integrated Database (eGRID2012)*, Data and Tools, Oct. 2015. [Online]. Available: https://www.epa.gov/energy/egrid.

[252] K. Siler-Evans, I. L. Azevedo, and M. G. Morgan, “Marginal emissions factors for the U.S. electricity system,” *Environmental Science & Technology*, vol. 46, no. 9, pp. 4742–4748, May 2012, ISSN: 0013-936X, 1520-5851. DOI: 10.1021/es300145v. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/es300145v.

[253] N. Z. Muller, “Toward the measurement of net economic welfare: air pollution damage in the U.S. national accounts—2002, 2005, 2008,” in *Measuring economic sustainability and progress*, ser. Studies in income and wealth volume 72, D. W. Jorgenson, J. S. Landefeld, and P. Schreyer, Eds., Chicago ; London: University of Chicago Press, 2014, pp. 429–459, ISBN: 978-0-226-12133-8.

[254] N. Z. Muller, R. Mendelsohn, and W. Nordhaus, “Environmental accounting for pollution in the United States economy,” *American Economic Review*, vol. 101, no. 5, pp. 1649–1675, Aug. 2011, ISSN: 0002-8282. DOI: 10.1257/aer.101.5.1649. [Online]. Available: http://pubs.aeaweb.org/doi/abs/10.1257/aer.101.5.1649.

[255] J. Heo, P. J. Adams, and H. O. Gao, “Reduced-form modeling of public health impacts of inorganic PM2.5 and precursor emissions,” *Atmospheric Environment*, vol. 137, pp. 80–89, Jul. 2016, ISSN: 13522310. DOI: 10.1016/j.atmosenv.2016.04.026. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1352231016303090.

[256] R. Martin, *Wall Street's Quest To Process Data At The Speed Of Light*, Apr. 2007. [Online]. Available: http://www.informationweek.com/wall-streets-quest-to-process-data-at-th/199200297.

[257] T. Kunath, *Preserving TelePresence Quality over the WAN with Performance Routing*, 2011. [Online]. Available: http://www.cisco.com/web/services/news/ts_newsletter/tech/chalktalk/archives/201104.html.

[258] T. Szigeti and C. Hattingh, "Quality of service design overview," in *End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs*, 1st, Cisco, 2005, p. 768, ISBN: 1-58705-176-1. [Online]. Available: http://www.ciscopress.com/articles/article.asp?p=357102&seqNum=2.

[259] I. Grigorik, *Latency: The New Web Performance Bottleneck*, Blog, Jul. 2012. [Online]. Available: https://www.igvita.com/2012/07/19/latency-the-new-web-performance-bottleneck/.

[260] Verizon, *IP Latency Statistics*, 2016. [Online]. Available: http://www.verizonenterprise.com/about/network/latency/.

[261] R. Krishnan, H. V. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao, "Moving beyond end-to-end path information to optimize CDN performance," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '09, New York, NY, USA: ACM, 2009, pp. 190–201, ISBN: 978-1-60558-771-4. DOI: 10.1145/1644893.1644917. [Online]. Available: http://doi.acm.org/10.1145/1644893.1644917.

[262] E. Schurman and J. Brutlag, *The User and Business Impact of Server Delays, Additional Bytes, and HTTP Chunking in Web Search*, San Jose, CA, Jun. 2009. [Online]. Available: http://conferences.oreilly.com/velocity/velocity2009/public/schedule/detail/8523.

[263] J. Brutlag, *Speed Matters*, Jun. 2009. [Online]. Available: http://googleresearch.blogspot.com/2009/06/speed-matters.html.

[264] J. Liddle, *Amazon found every 100ms of latency cost them 1% in sales*, Blog, Aug. 2008. [Online]. Available: http://blog.gigaspaces.com/amazon-found-every-100ms-of-latency-cost-them-1-in-sales/.

[265] S. Work, *How Loading Time Affects Your Bottom Line*, 2011. [Online]. Available: https://blog.kissmetrics.com/loading-time/.

[266] M. Prince, *The Relative Cost of Bandwidth Around the World*, Aug. 2014. [Online]. Available: http://blog.cloudflare.com/the-relative-cost-of-bandwidth-around-the-world/.

[267] Colocation America, *Data Center Bandwidth and Costs*, Aug. 2015. [Online]. Available: http://www.colocationamerica.com/data-center-connectivity/bandwidth.htm.

[268] M. Bolinger and J. Seel, "Utility-scale solar 2014: an empirical analysis of project cost, performance, and pricing trends in the United States," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-1000917, Sep. 2015. [Online]. Available: https://emp.lbl.gov/sites/all/files/lbnl-1000917_0.pdf.

[269] R. H. Wiser and M. Bolinger, "2014 Wind technologies market report," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-188167, Aug. 2015. [Online]. Available: https://emp.lbl.gov/sites/all/files/lbnl-188167_0.pdf.

[270] U.S. Energy Information Administration, "Levelized cost and levelized avoided cost of new generation resources in the Annual Energy Outlook 2015," Washington, D.C., Tech. Rep., Jun. 2015. [Online]. Available: https://www.eia.gov/forecasts/aeo/pdf/electricity_generation.pdf.

[271] Lazard, "Levelized Cost of Energy Analysis," Tech. Rep. version 8.0, Sep. 2014. [Online]. Available: https://www.lazard.com/media/1777/levelized_cost_of_energy_-_version_80.pdf.

[272] Akamai, *Facts & Figures*, 2016. [Online]. Available: https://www.akamai.com/us/en/about/facts-figures.jsp.

[273] Akamai Technologies Inc, *Akamai reports fourth quarter 2015 and full-year 2015 financial results*, Feb. 2016. [Online]. Available: http://www.prnewswire.com/news-releases/akamai-reports-fourth-quarter-2015-and-full-year-2015-financial-results-300217658.html.

[274] North Carolina Department of Revenue, *Important Notice: Qualifying Datacenter*, Dec. 2015. [Online]. Available: http://dornc.com/taxes/sales/impnotice121515_datacenter.pdf.

[275] N. Z. Muller, “Linking policy to statistical uncertainty in air pollution damages,” *The BE Journal of Economic Analysis & Policy*, vol. 11, no. 1, 2011. [Online]. Available: http://www.degruyter.com/view/j/bejeap.2011.11.issue-1/bejeap.2011.11.1.2925/bejeap.2011.11.1.2925.xml.

[276] D. G. Ware, “February weather records broken in N.Y.C., elsewhere,” *UPI*, Feb. 2015. [Online]. Available: http://www.upi.com/Top_News/US/2015/02/28/February-weather-records-broken-in-NYC-elsewhere/5521425098804/.

[277] J. S. Graff Zivin, M. J. Kotchen, and E. T. Mansur, “Spatial and temporal heterogeneity of marginal emissions: implications for electric cars and other electricity-shifting policies,” *Journal of Economic Behavior & Organization*, vol. 107, pp. 248–268, Nov. 2014, ISSN: 01672681. DOI: 10.1016/j.jebo.2014.03.010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167268114000808.

[278] K. Siler-Evans, I. L. Azevedo, M. G. Morgan, and J. Apt, “Regional variations in the health, environmental, and climate benefits of wind and solar generation,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 29, pp. 11 768–11 773, Jul. 2013, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1221978110. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1221978110.

Appendix A

Marginal Damage Factor Estimation

Marginal damage factors (MDFs) for the U.S. electricity grid are used in Chapter 4. This appendix provides additional information on how these estimates are derived. The general approach is to obtain hourly emissions and fossil generation data at the plant level [249], convert these emissions to damages using the AP2 [253] and EASIUR [255] emissions damage models for the criteria pollutants and a social cost of carbon of \$40/ton for CO_2, aggregate hourly damages at a geographic region, and use regression to estimate regional MDFs (i.e., \$/MWh). The geographic regions used are the eGRID subregions, with the three New York subregions (NYCW, NYLI, and NYUP) grouped into a new subregion called NWYK.

A.1 Relationship of damages to fossil generation

We began with a conceptual model of the emissions damages[1] generating process. Fossil-fueled power plants generate emissions, and this emissions rate per MWh is not constant: it changes depending on the load of each individual power plant. As overall regional load increases, different types of plants come on line, which changes the emissions rate: if a gas peaker ramps up in a region where the base load is hydro and nuclear, the running average regional emissions rate will increase, while the same peaker in a region where the base load is coal will cause the running average to decrease. (This varying emissions rate is the argument against using simple average emissions factors.)

Figure A.1, which shows the electricity dispatch curve for the Southeast in summer 2010 (top) and 2012 (bottom), respectively, illustrate this concept. For 2010, we would expect the MDF curve be very low at the bottom, increasing as load-following coal comes online, decreasing when natural gas becomes the marginal fuel, and finally increasing again when petroleum peakers are required. In 2012, cheap natural

[1] In this example, power plant emissions are translated into power plant damages, as discussed elsewhere in this report. One might think of the power plant emitting dollars instead of pollutants in this case. However, this same process has been used to generate marginal *emissions* factors instead of marginal *damage* factors.

gas prices have caused that fuel dispatch earlier than coal in some cases, so the MDF curve will probably be less smooth as the marginal plant alternates between these two fuels.

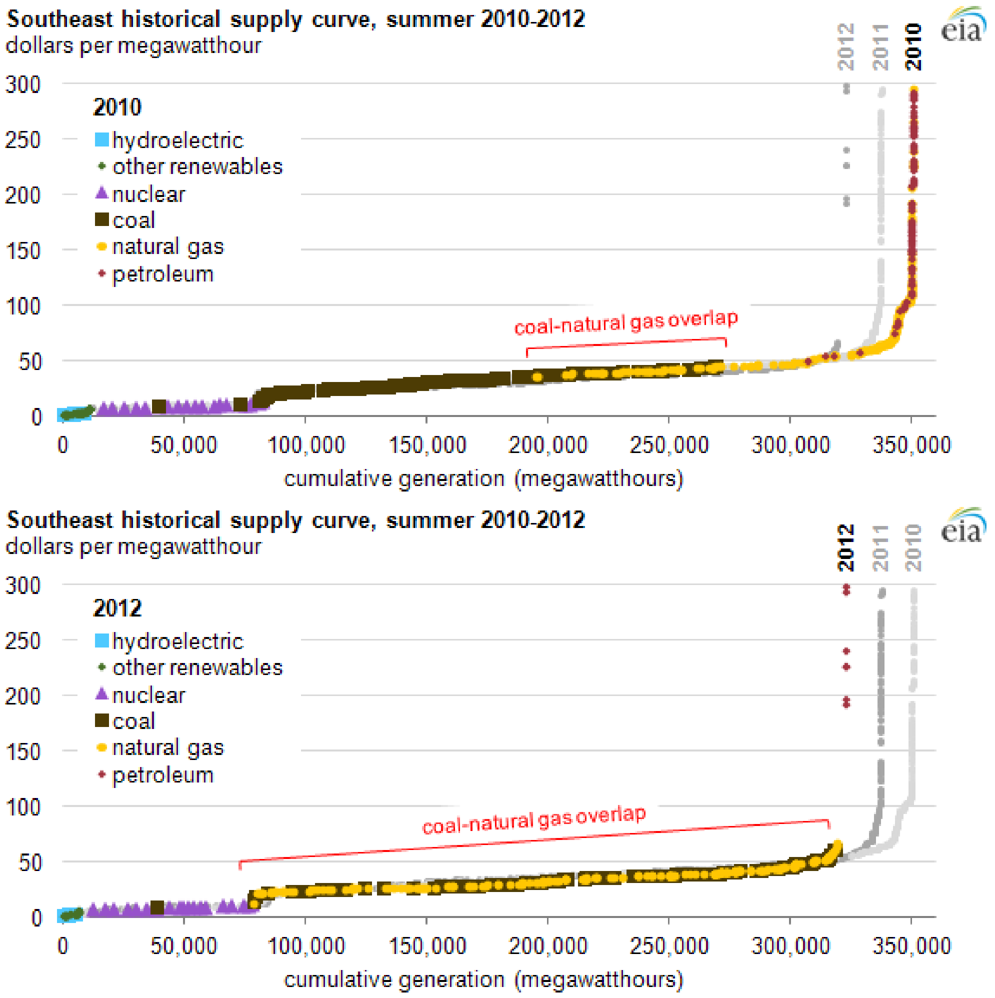


Figure A.1: Electricity dispatch curve for the Southeast, summer 2010 (top) & 2012 (bottom). Source: U.S. government work (EIA) / Public domain.

With this model in mind, we now investigate the relationship of damages to fossil generation in our data. Figures A.2 – A.5 show plots of damages against fossil generation. In general, as expected, each unit of electricity produced results in harmful emissions, so as generation increases, damages increase. The notable exception is NO_x in New York, where these emissions have a *benefit* in the AP2 model. The explanation for this counterintuitive finding is that, in dense urban areas, NO_x can reduce ozone levels, resulting in a net benefit [275]. AP2 estimates negative damages for NO_x in New York and parts of California. EASIUR does not estimate any net negative damages.

We also observe that, in some cases, the scatterplot appears to indicate two different slopes (e.g., SO_2

in NEWE and NWYK). Further investigation of New York, where this observation is the most marked, reveals that this pattern shows up in the untransformed emissions plots (i.e., before converting to damages), and that the steeper-sloped observations occur roughly around January 7-8, and the 2nd half of February. These time periods corresponded with an unusually brutal winter weather pattern, with New York City seeing the coldest February in 81 years, and Boston seeing near-record snowfalls [276]. An analysis of the interaction between SO_2 emissions and cold weather is beyond the scope of this work, but it seems likely that the spike in SO_2 is related to the weather pattern; the fact that this dual-slope clustering is most evident in the northern regions supports this conclusion. This type of event represents an opportunity where data center load shifting may deliver much larger public benefits than normal.

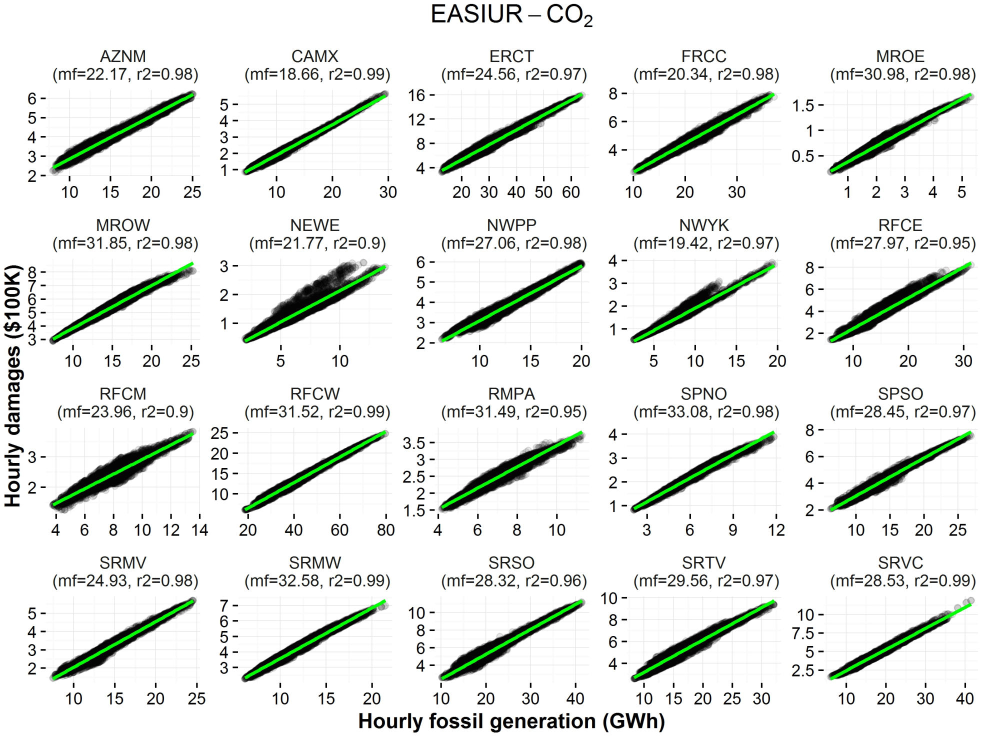


Figure A.2: Regression of CO_2 damages on fossil generation. *Note varying x- and y-axis scales. This plot can be used to see the distribution of observations and assess fit of the linear model, but it should not be used to visually compare the slopes among the regions.*

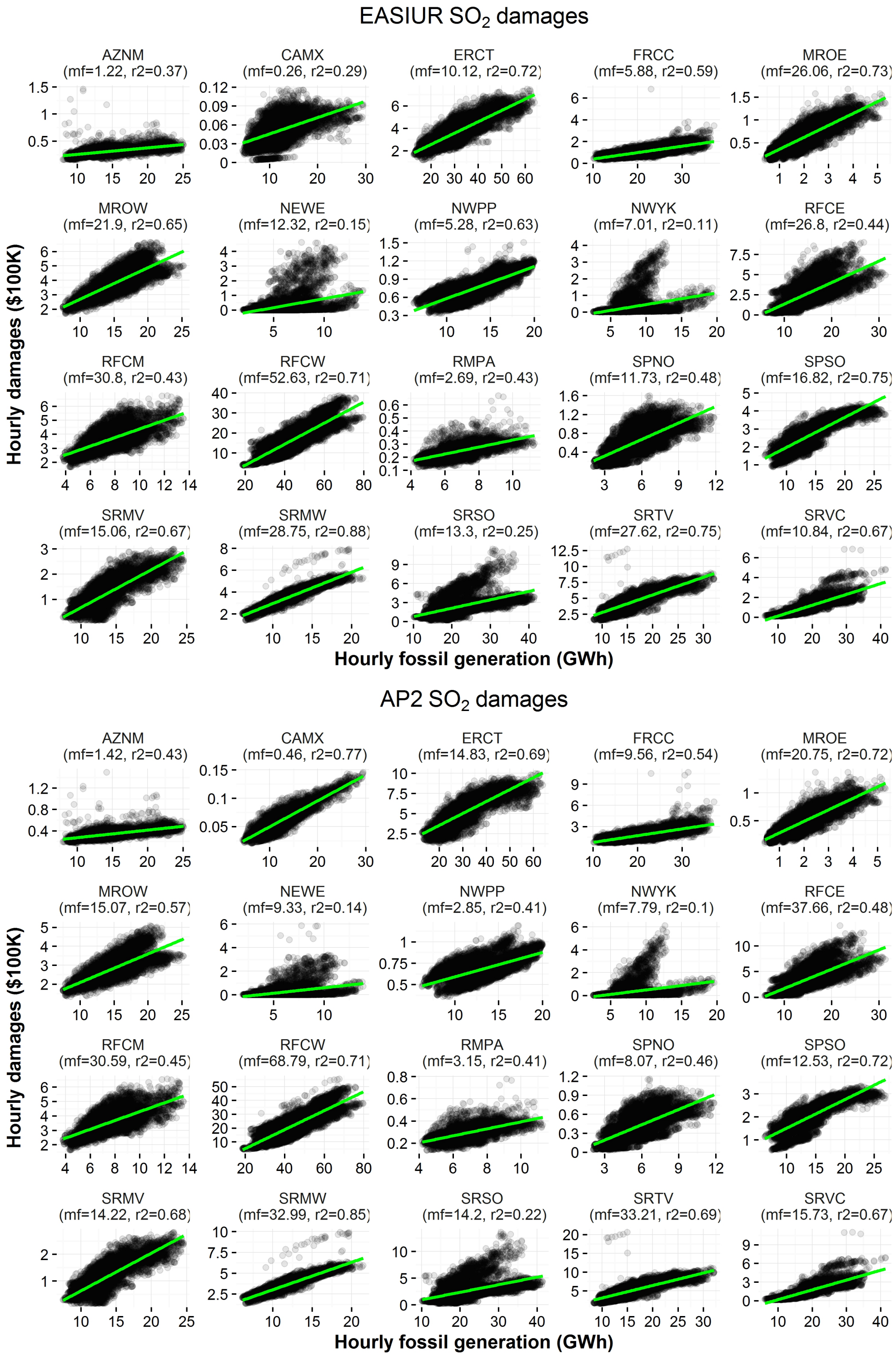


Figure A.3: Regression of SO_2 damages on fossil generation, EASIUR & AP2. *Note varying x- and y-axis scales. This plot can be used to see the distribution of observations and assess fit of the linear model, but it should not be used to visually compare the slopes among the regions.*

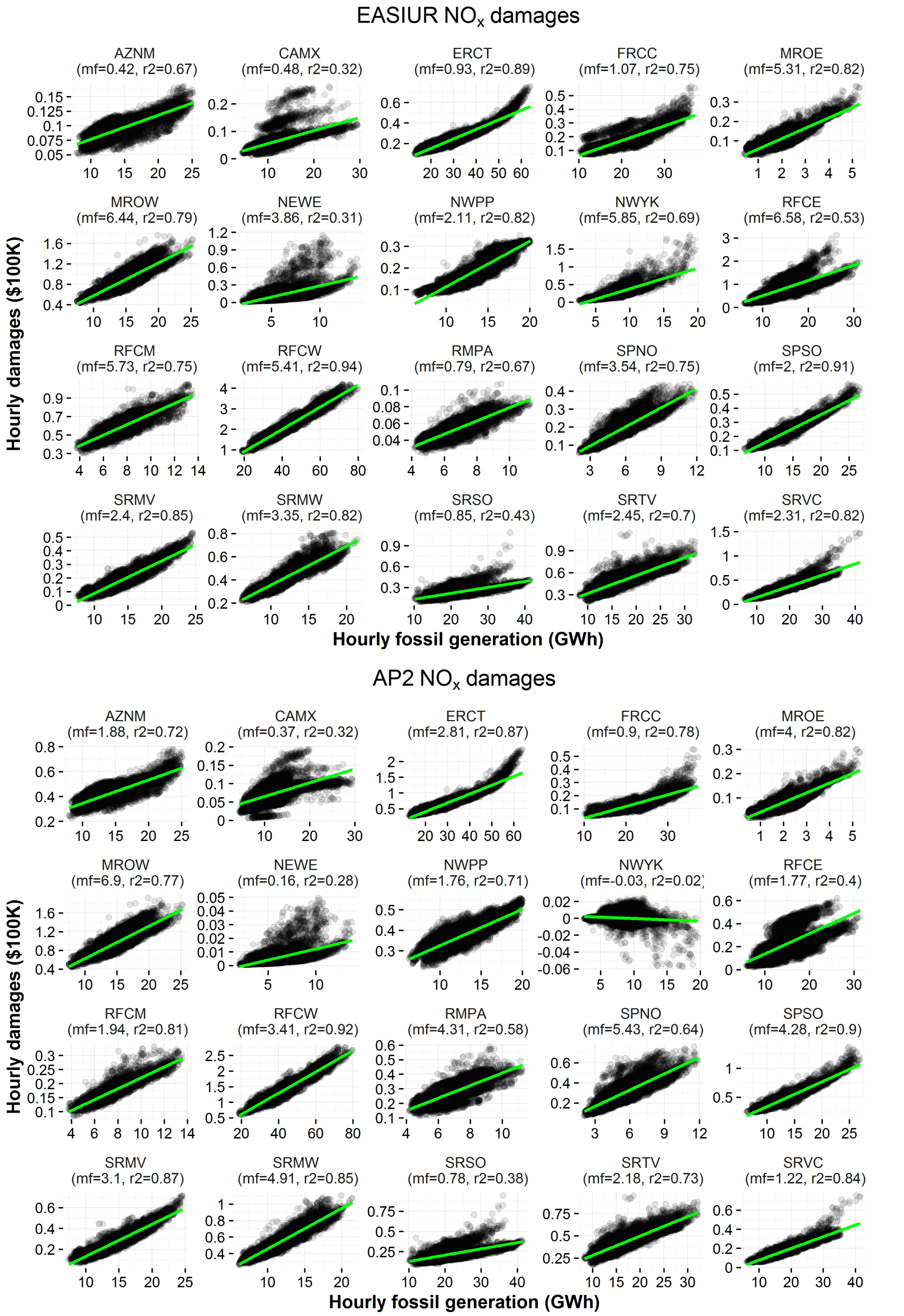


Figure A.4: Regression of NO$_x$ damages on fossil generation, EASIUR & AP2. *Note varying x- and y-axis scales. This plot can be used to see the distribution of observations and assess fit of the linear model, but it should not be used to visually compare the slopes among the regions.*

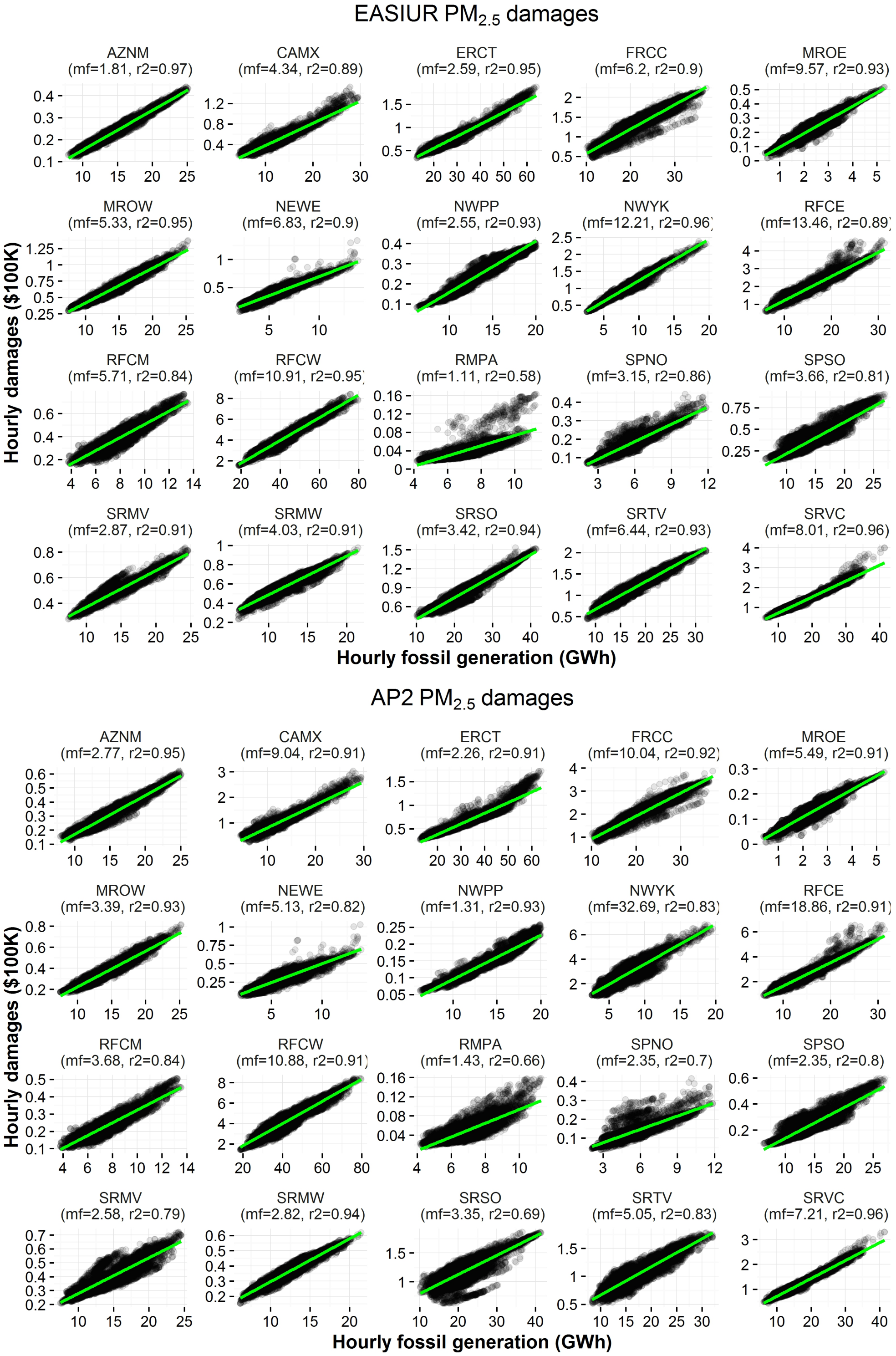


Figure A.5: Regression of PM$_{2.5}$ damages on fossil generation, EASIUR & AP2. *Note varying x- and y-axis scales. This plot can be used to see the distribution of observations and assess fit of the linear model, but it should not be used to visually compare the slopes among the regions.*

A.2 Estimating marginal damage factors using regression

By inspection of Figures A.2 – A.5, we can observe qualitatively that a linear model works reasonably well on the pooled data in many cases (e.g., for CO_2 in most regions), but not particularly well in others. The following example illustrates the use of regression to quantify the relationship between damages and fossil generation as a set of MDF estimates.

We instantiate the conceptual model described at the beginning of this appendix—that the emissions rate varies with generation load—in a regression model of the form:

$$D = \beta_0 + \beta_1 G + \epsilon \tag{A.1}$$

where D is damages, G is fossil generation, and β_1 estimates the MDF—that is, the expected change in damages for an incremental change in generation. However, as discussed above, we expect β_1 to vary over the range of generation load. Figure A.6, which shows the relationship between SO_2 damages and generation for the New England subregion (NEWE), serves as an illustrative example. We mentioned the impact of an extremely cold winter above; the record-setting chill in late February corresponds generally to the observations in Area A of the figure. We can also see that, at the upper range of generation (Area B), the damage factor appears to be greater than in the main body of observations, which is consistent with our understanding that generators high on the dispatch curve for this region are dirtier than plants lower on the curve. For these reasons, we move beyond a single linear fit to these data.

One way to improve the model is to add regressors. For instance, we have discussed evidence that temperature has a correlation with SO_2 damages—but there are likely two mechanisms for this correlation: emissions seem to increase in response to temperature irrespective of generation (Area A), but colder periods might also have higher load due to increased space heating, which should also increase SO_2 (Area B). Adding temperature as a variable in the regression, possibly with an interaction term, might disentangle these effects and would almost certainly improve the fit of the model. However, the focus of this work being on application of MDFs rather than on the generation of new models for estimating these factors, we necessarily limit ourselves to the sort of approaches used in Siler-Evans *et al.* [252] and Graff Zivin *et al.* [277], which do not use temperature. Part of the appeal of these approaches is that they allow reasonable estimation of emissions and damage factors with simple models, and the results can

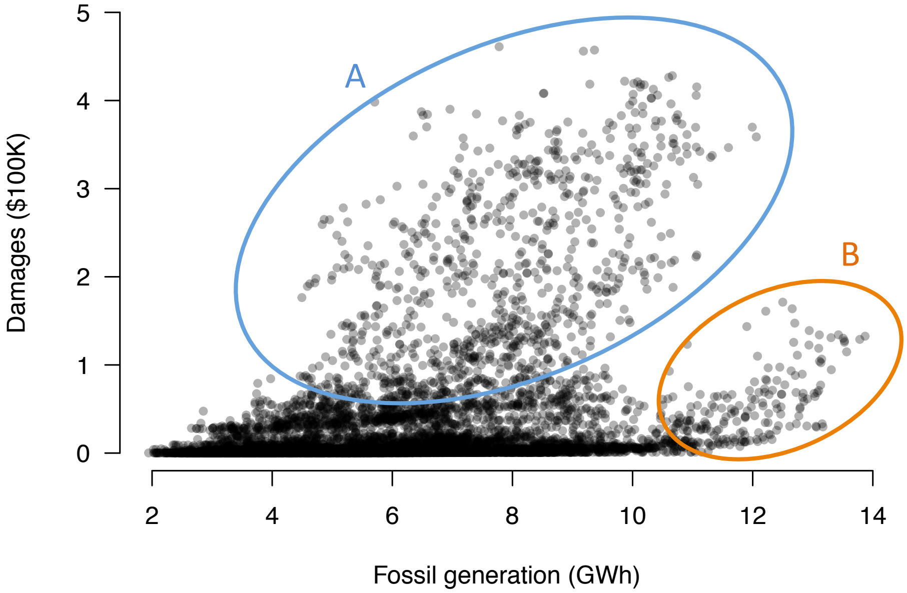


Figure A.6: SO_2 damages vs. fossil generation for New England (NEWE). Area A shows observations during cold-weather event of February, 2015. Area B shows high-load observations.

be broadly applied. A model that included temperature would require any application of the MDFs to estimate the temperature input. However, we will introduce a partitioning scheme that includes a rough proxy for temperature below.

A.2.1 Hourly partitioning as a proxy for demand

First, we concentrate on handling the nonlinearity in the model evidenced by Area B in Figure A.6, while noting that in certain regions (such as in the Northeast) we may have to accept larger error terms in exchange for simplicity. The nonlinearity arises from the fact that, as the marginal plant changes, the marginal fuel type changes, and thus the marginal damage factor changes. There is little point in estimating marginal effects with the linear pooled model shown above, since the result is a "marginal" damage factor that remains constant over the range of the dispatch curve. One way of differentiating the MDFs for different areas of the dispatch curve is to partition it, estimating different coefficients each partition. (We could also explore variable transformations, which might allow a better fit at the expense of interpretative simplicity.)

First, we partition for each hour of the day, since load follows a generally predicable diurnal pattern (Figure A.7), and estimates a different MDF for each hour of the day in each region. The results of this model (using damages from EASIUR) for NEWE are shown in Figure A.8. We can observe that there is some variation in the slope estimates from hour to hour, and that the r-squared statistics are a little bit better than the corresponding model in Figure A.3, but also that cold-temperature observations (Area A) are not explained in the regression. That is, hour of day does not completely explain differences in MDF, which is expected, since we expect temperature to have an effect on load as well.

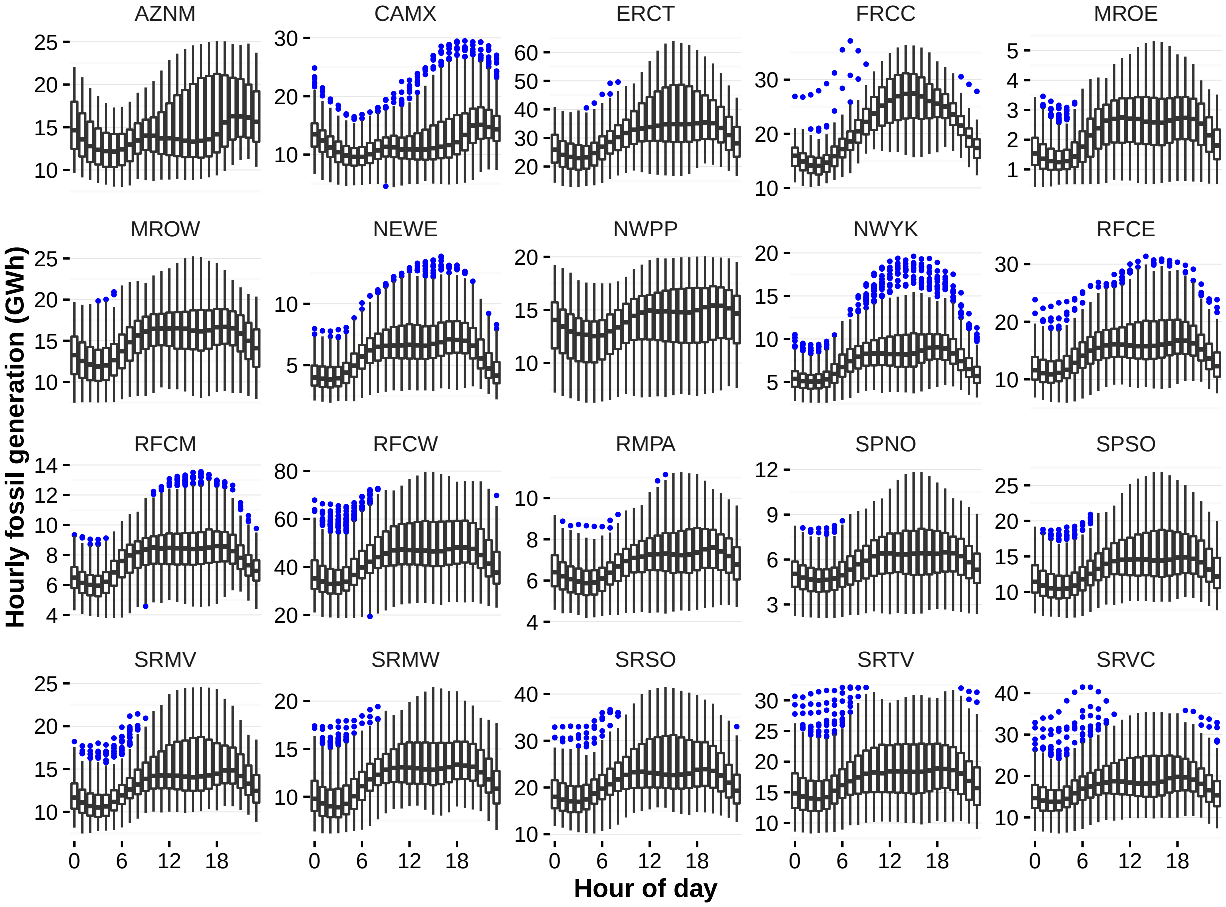


Figure A.7: Boxplot of hourly fossil generation by region showing diurnal pattern. Note differing y-axis scales. The hour of day is normalized to Eastern time, so the load profile is shifted later for regions in the Central (1 hour), Mountain (2 hours), and Pacific (3 hours) time zones. Box upper and lower limits correspond to the interquartile range (IQR); whiskers extend to $1.5 * IQR$ beyond the whiskers; observations beyond whiskers are plotted as blue points.

A.2.2 Seasonal partitioning as a proxy for temperature

We can expand this approach by differentiating by season as a proxy for temperature to the model. A regression in which the data are partitioned not only on hour but also on season might perform better.

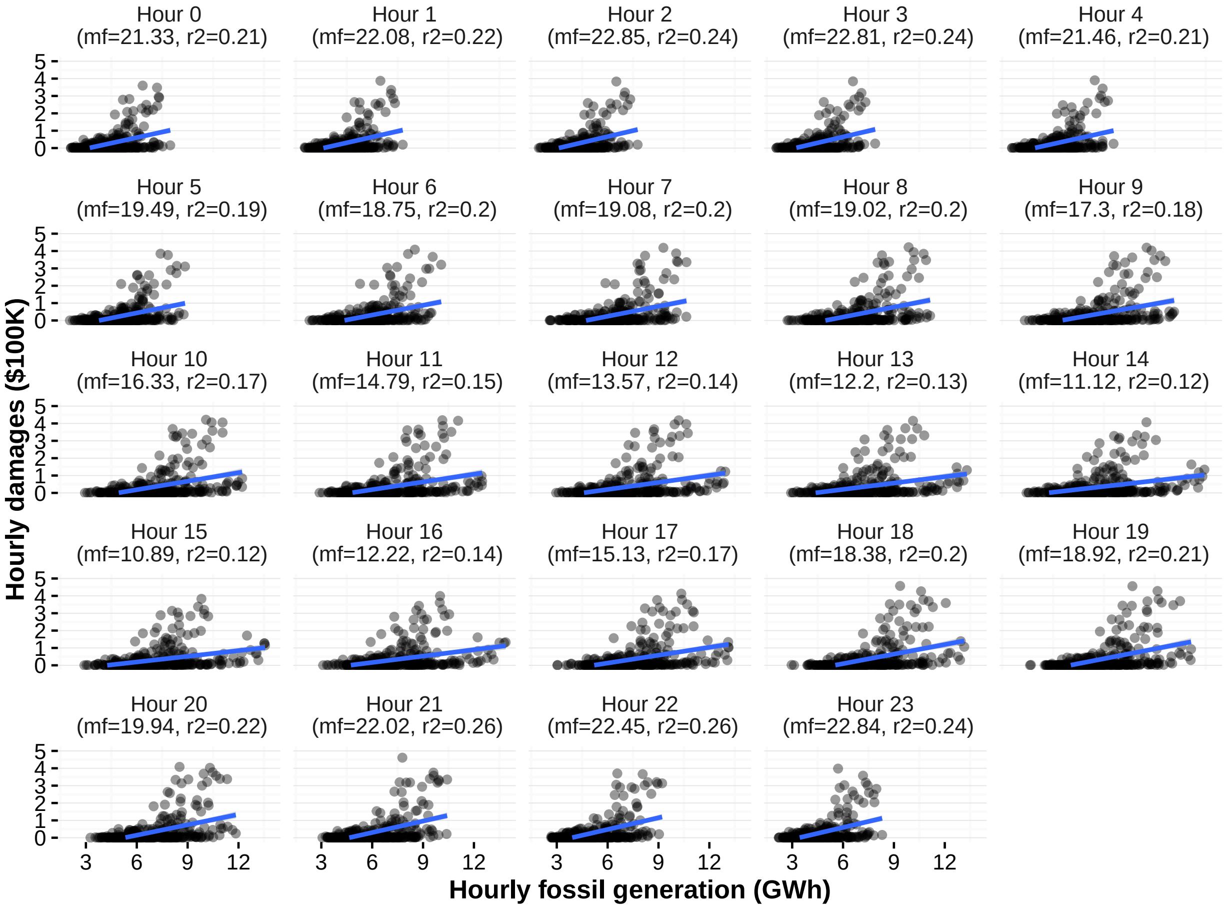


Figure A.8: SO_2 damages vs. fossil generation by hour for New England (NEWE).

These results are shown in Figure A.9. The MDF estimates are much larger than in the non-seasonal model, and the r-squareds have improved as well. The linear fit is still not especially good—the relationship of damage to generation still appears to be nonlinear over the entire generation range, with the higher range of generation appearing require a steeper slope. That is, in general terms we have addressed the Area A issue but are still confronted with Area B. However, the model gives us a ballpark estimate of the MDF.

Note also that in looking at SO_2 in NEWE we have chosen a particularly difficult subset of the data. The seasonal, hourly linear model achieves better performance for many of the other pollutants and regions, which are simpler to fit. Figure A.10, for example, shows the SO_2 hourly damages for the summer load in the Carolinas and Virginia. In general, the CO_2 regressions provide a very good fit, the $PM_{2.5}$ also reasonably good, with NO_x and SO_2 being more variable.

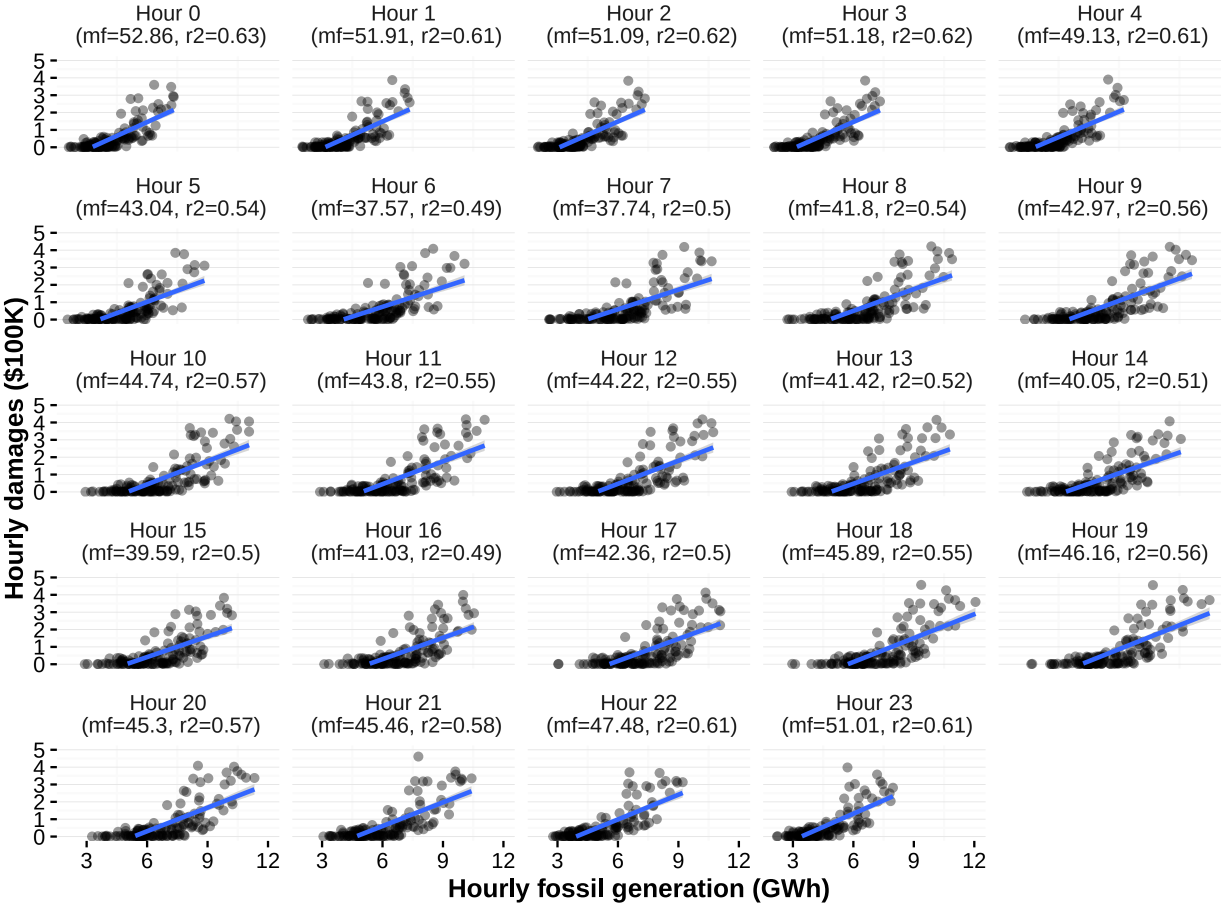


Figure A.9: SO_2 damages vs. fossil generation by hour for NEWE (New England), winter only.

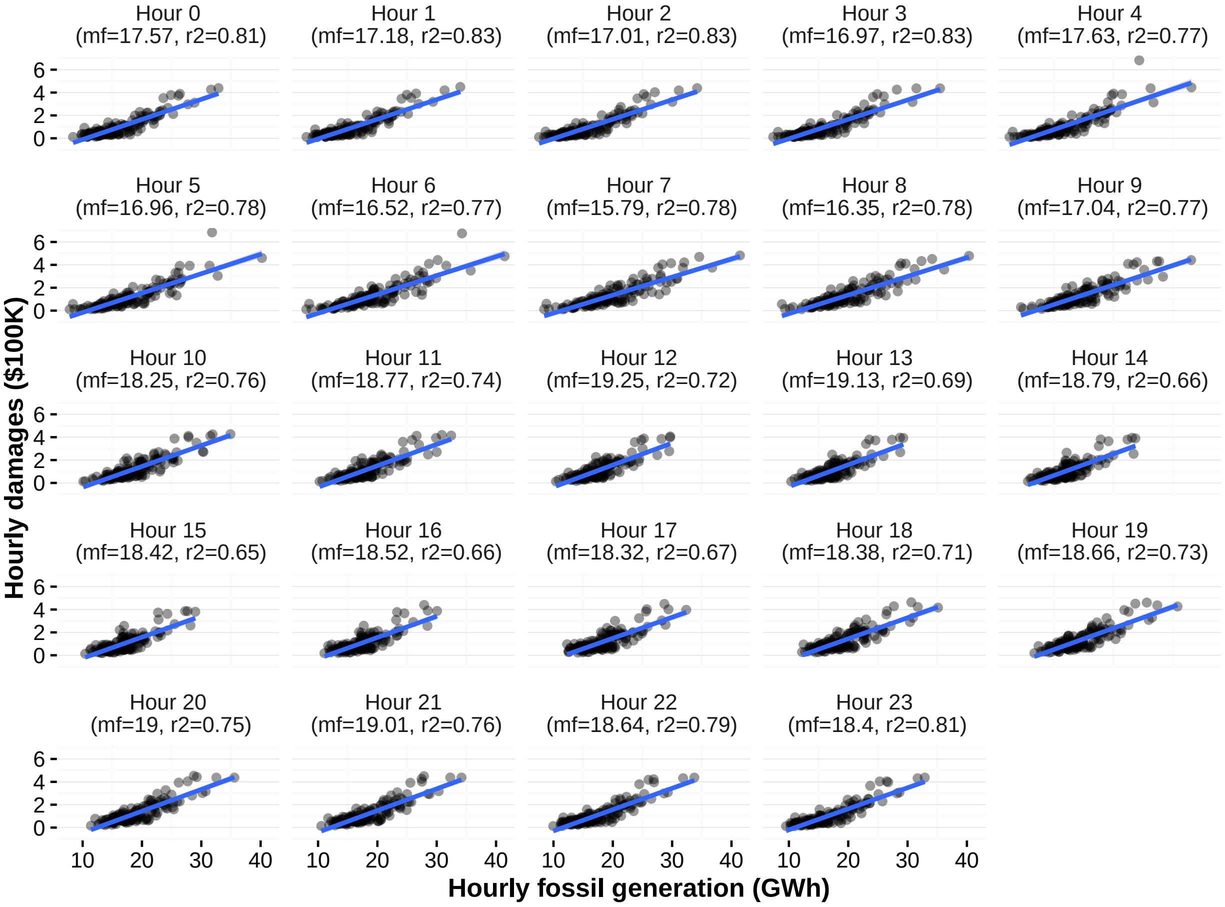


Figure A.10: SO_2 damages vs. fossil generation by hour for SRVC (Carolinas and Virginia), summer only.

A.3 Implicit assumptions, and attempts to obviate them

The regression approach described above implies two key assumptions. First, our generation data is aggregated from fossil plants only and thus excludes nuclear and renewable generation sources, so these MEFs assume that the marginal generator is coal-, gas-, or oil-fired. Second, this approach does not account for imports and exports between regions. Therefore, we assume that demand in a particular subregion is met by generation in the same subregion. This assumption is the reason we calculate MEFs at the subregion rather than the state level; using larger geographic areas mitigate somewhat the import-export concern.

We have explored other approaches to remove the need for these assumptions, including using hourly *demand*, obtained from from FERC Form 714 reports [221], to address the fossil-only issue, and including regressors for generation in neighboring regions to account for imports. However, doing so raises a few additional difficulties. The FERC demand data is reported at the BA level, and, as noted above, BA boundaries do not line up with the regions and subregions used by the EPA. (See Figures 4.5 and 4.6.) Thus, allocating demand to the emissions regions (let alone subregions or states[2]) is difficult to do without relying on gross assumptions that would likely not be accurate. Areas in the MISO footprint, which spans parts of four NERC regions (MRO, SERC, SPP, and RFC), are particularly problematic.

Additionally, there is some double-counting in the FERC data; some smaller BAs report demand along with the larger BAs into which they have been consolidated. We can make informed decisions about which BAs should likely be dropped, and we can conduct a rough check on those guesses by comparing aggregate demand in the retained BAs against regional demand reported by the EIA. However, these two issues—allocation and double-counting—result in a large amount of data uncertainty. Perhaps the most promising attempt to address these issues and use demand data is Graff Zivin *et al.* [277], although they, too, are limited in fidelity and focus on NERC regions while not allowing imports and exports among the three U.S. interconnects.

The choice of approach comes down to using fairly reliable data with some limiting assumptions, or higher-uncertainty data in an attempt to develop a more realistic model. So far, this approach has not yielded estimates in which we are confident enough to use. Thus, for this work, we elect the approach described in this appendix; assumptions notwithstanding, we are more confident in the validity of its

[2] In the words of my colleague, Kyle Siler-Evans, with whom I have worked on methods for estimating marginal damages: "At some point when going to higher resolution, you start introducing more error than you are resolving."

results, and it has already been used to analyze regional impacts of energy interventions [e.g., 278]. (An additional mark against the second approach is that FERC demand data are, as of the time of this writing, not yet available for 2015, the year for which we have price and traffic data.)

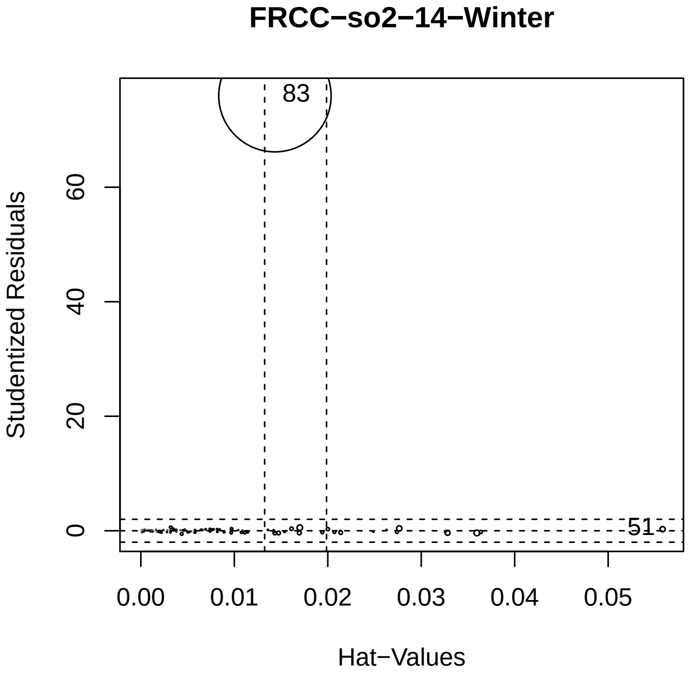
A.4 Differenced model

Siler-Evans *et al.* [252] uses a differenced model for estimating MDFs, in which the general formula is:

$$\Delta E = \beta \Delta G + \epsilon \tag{A.2}$$

In exploring this regression formula with our data, we found that the results were very similar to the non-differenced regression in "well-behaved" cases such as most of the CO_2 observations. However, for other pollutants, the results were less stable. A particular problem is that differencing appears to exacerbate the effect of high-leverage outliers. For example, Figure A.11 shows influence plots for Hours 14 and 15 for the differenced wintertime SO_2 regression for FRCC, with damages estimated by EASIUR. Observations with both high discrepancy (i.e., difference from the mean) and high leverage (i.e., outside the normal range of the data) have high influence on the regression line. In this figure, each observation is plotted against discrepancy (vertical axis) and leverage (horizontal axis). The horizontal dashed lines represent $+/-$ 2 standard deviations on the residual scale, and the vertical dashed lines represent 2x and 3x the average leverage of the data. The area of the plotted point corresponds to its Cook's distance, a measure of influence. Each observation in this case corresponds to a particular day of the winter season, and Observation 83 has extremely large residual and also has relatively high leverage, resulting in extraordinary influence on the slope of the regression line.

Observation 83 corresponds to March 24, 2015. Table A.1 shows fossil generation and SO_2 damages from 9:00 – 18:00 on this date. We can see that the damages (i.e., the emissions) of SO_2 jumped sevenfold in Hour 14 and then declined back to normal levels in Hour 15. Fossil generation was in the midst of a gradual peaking and was relatively high as well, but by no means saw so great a jump. This anomaly is responsible for the large influence seen for this observation. With Observation 83, the coefficient for Hour 15 is *negative* \$16.9/MWh, wheras we know that the slope should be positive. This one observation has flipped the sign of the coefficient.



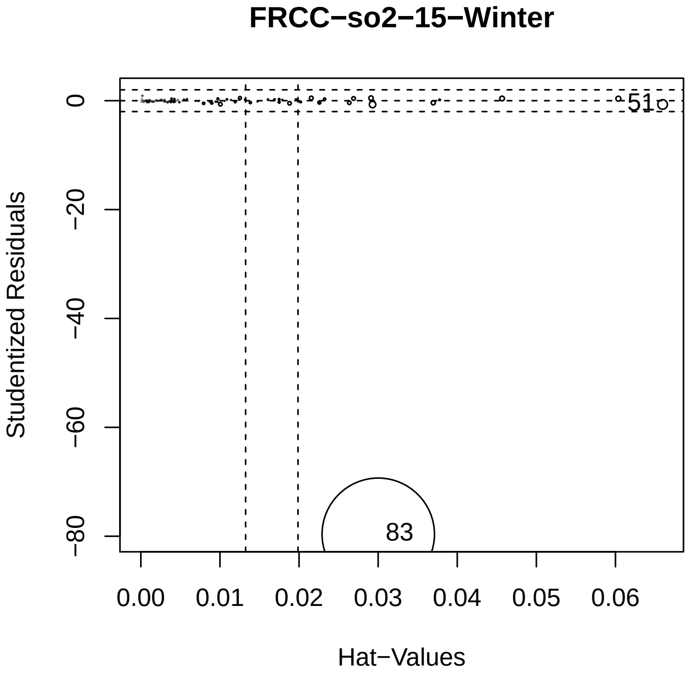


Figure A.11: Influence plot for differenced model of SO_2 MDF for FRCC Winter Hours 14 & 15 shows very high influence of observation 83 on the regression slope. The horizontal dashed lines represent +/− 2 standard deviations on the residual scale, the vertical dashed lines represent 2x and 3x the average leverage of the data, and the area of each plotted point corresponds to its Cook's distance. EASIUR is the damage model used.

Table A.1: Hourly fossil generation and SO_2 damages in FRCC, March 24, 2015.

Hour	Fossil Generation (GWh)	SO_2 Damages ($thousands)
9	19.3	77.2
10	20.2	77.7
11	20.8	80.3
12	21.5	86.1
13	22.3	106.1
14	23.0	**681.5**
15	23.7	109.0
16	23.9	92.1
17	23.3	82.8
18	22.9	78.0

It is tempting to just eliminate this outlier as an obvious anomaly—either it is a truly erroneous observation, resulting perhaps from a sensor malfunction or data recording error, or it depicts a low-probability event such as backup generator response resulting from an unexpected plant shutdown. In either case, one could argue, it should be removed from our estimates of "normal" damage factors. In this case, the case for removal is probably compelling. The issue, however, is that there is no clear demarcation for when a point should probably be removed. Indeed, automated screening based on a Cook's distance threshold of 1[3] identified over 500 data points with outsized influence on their regressions. In most cases, the influence is not as extreme in this example, and the decision to remove the point is even more subjective. We might attempt to address this problem by manually reviewing these points and labeling them with a dummy variable, to be included in the model.

The non-differenced model seems more robust to these types of data points. In comparison, less than 100 observations were flagged for high influence. Figure A.12 shows the influence plot the SO_2 damages in FRCC for wintertime hour 15 for the standard (non-differenced) regression. None of the observations have inordinately high influence, and none were flagged in our screening process, although we might be tempted to investigate those points near the plot limits.

While we believe there is promise in refining the differenced version of the model, for this work we opt for the more straightforward and less volatile standard linear regression model. Clearly there is much room for improvement: we might attempt adding more variables to the model or trying variable transformations to represent linearize nonlinear relationships. These modifications, of course, come at a cost of simplicity.

[3] An arbitrary, if commonly used, threshold.

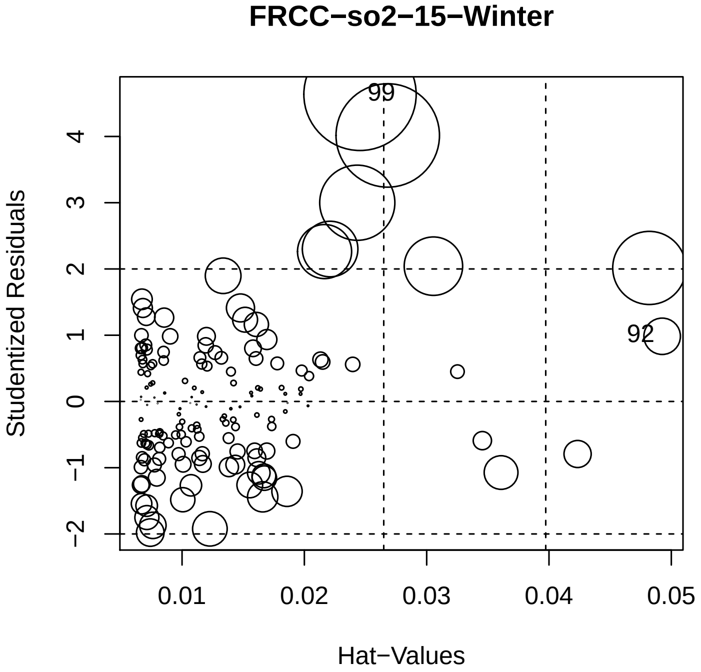


Figure A.12: Influence plot for standard model of SO_2 MDF for FRCC Winter Hours 15 shows no obviously problematic observations. Compare with Figure A.11, bottom. The horizontal dashed lines represent +/− 2 standard deviations on the residual scale, the vertical dashed lines represent 2x and 3x the average leverage of the data, and the area of each plotted point corresponds to its Cook's distance. EASIUR is the damage model used.

A.5 Summary of damage factor estimates used in this analysis

Ultimately, it is critically important to understand the origins and limitations of these (and, in fact, any) MDF estimates. In this particular case, our MDFs are subject to data uncertainty, adoption of limiting assumptions, model error, and uncertainty in the damage models, as documented in the previous sections. Despite these sources of uncertainty, these estimates provide a useful alternative to average emissions factors in examining the impact of interventions that affect electricity consumption.

This section summarizes the MDF estimates used in the load shifting analysis in Chapter 4. Figures A.13 to A.16 compare the seasonal hourly marginal and average damage factors using both EASIUR and AP2 damage values by pollutant and subregion. The factor estimates show inter-regional differences that are functions of climate, fuel mix, and system demand factors such as prevalence of air conditioning or electric heating. A detailed breakdown of these estimates is beyond the scope of this work, but we briefly highlight the sorts of general conclusions that can be drawn from the plots.

Looking at SO_2, we observe that, while either AP2 or EASIUR generate consistently higher damage factors within a region, this consistency does not extend between subregions. (I.e., in some subregions, the AP2 factors are larger, and in others, the EASIUR factors are larger.) In most regions, the winter factors are higher than those in the summer; this is particularly evident in subregions with cold climates such as NEWE and NWYK. In some regions (e.g., SRMV and NEWE), the marginal factors are consistently higher than the average factors; in others (e.g., RMPA) the reverse is true. We would expect, in general, that average factors might be higher than marginal factors where the peaker plants are comparatively cleaner than the base load, so that as load increases, the marginal emissions factor declines.

Finally, the diurnal patterns in each subregion can provide some insight into the fuel mix and dispatch curve. SRSO, for instance, generates about a third of its electricity from nuclear and hydro, a third from coal, and a third from natural gas. Observing how the winter MDFs change over the course of the day, we might posit that as load increases during the morning, coal replaces the clean base load as the marginal fuel on the dispatch curve, while natural gas peakers come online in the early afternoon, reducing the MDF. SPSO, in contrast, has a fuel mix of roughly 40% coal, 40% natural gas, and 20% wind and hydro. Wind variability makes the picture a little less clear, but we do observe a pronounced afternoon trough during the summer season, as would be expected when natural gas displaces coal as the marginal fuel to meet space cooling demand.

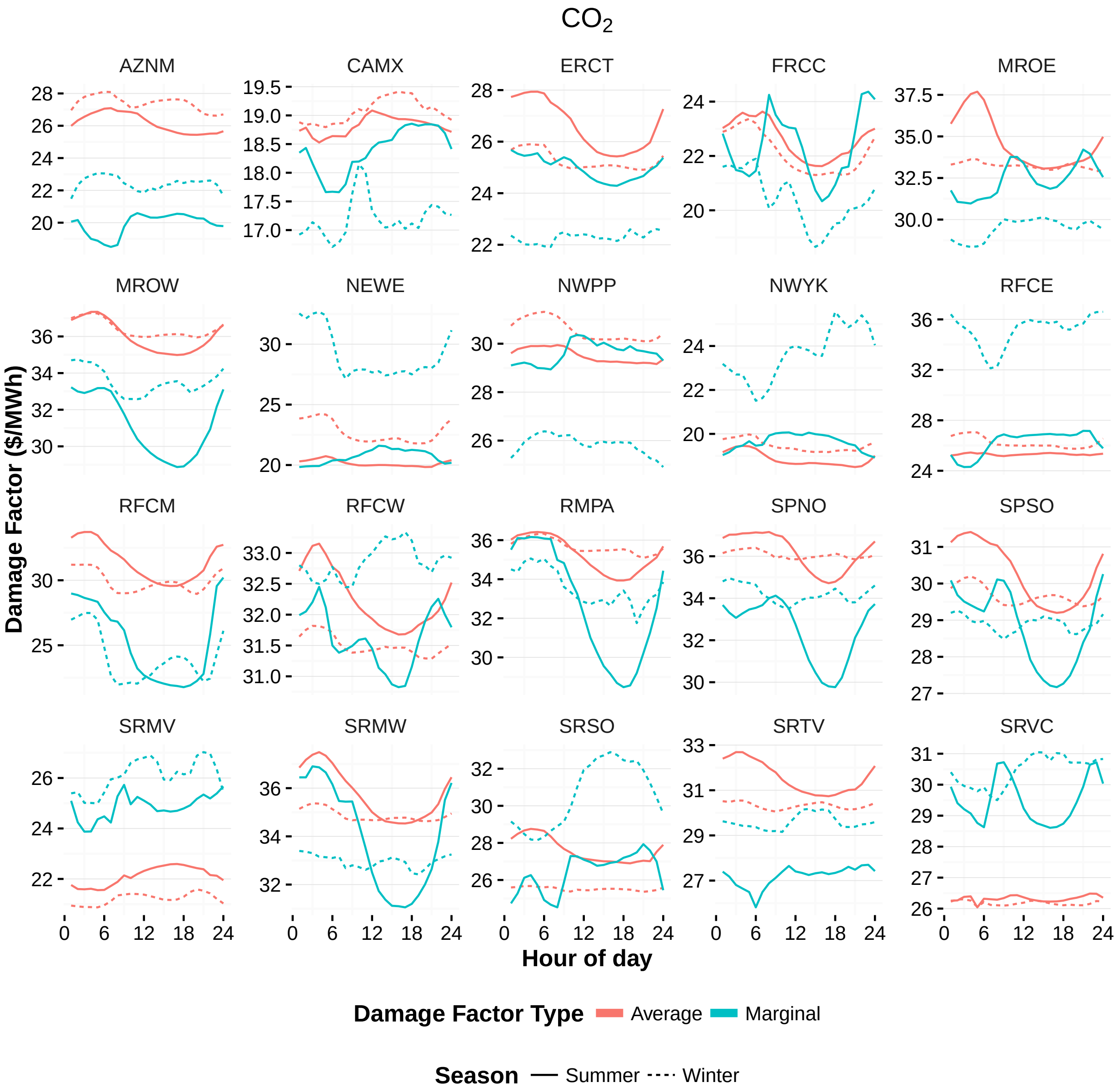


Figure A.13: CO_2 seasonal hourly marginal and average damage factors by subregion. Note that the **y-axis scale varies and does not start at zero** to aid visualization of the patterns among the various damage factor estimates within each subregion. See Figure A.36 for an accurate inter-regional comparison, and Figures A.17 – A.35 for a full-scale representation of the damage estimates within each region.

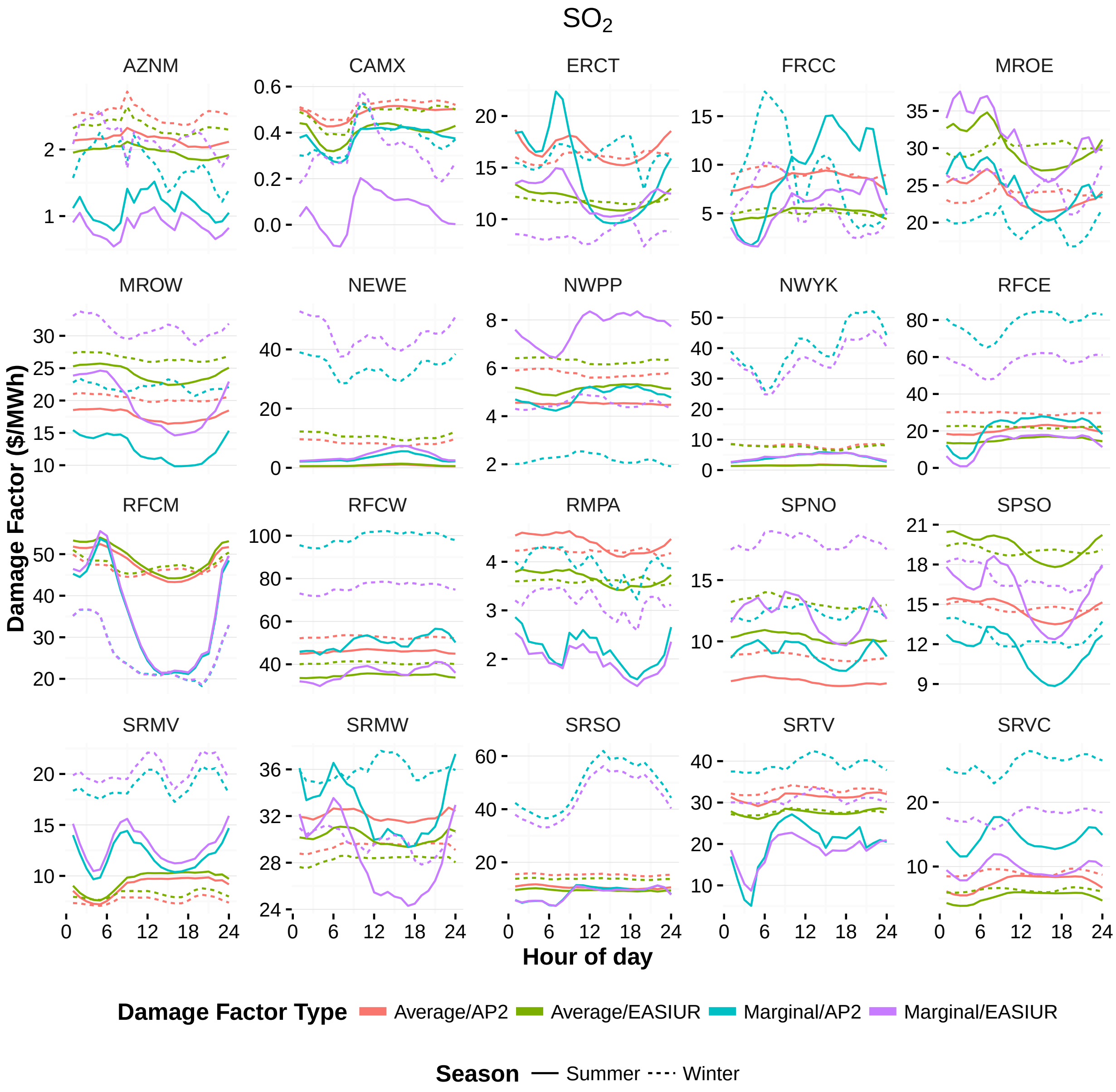


Figure A.14: SO_2 seasonal hourly marginal and average damage factors by subregion. Note that the **y-axis scale varies and does not start at zero** to aid visualization of the patterns among the various damage factor estimates within each subregion. See Figure A.36 for an accurate interregional comparison, and Figures A.17 – A.35 for a full-scale representation of the damage estimates within each region.

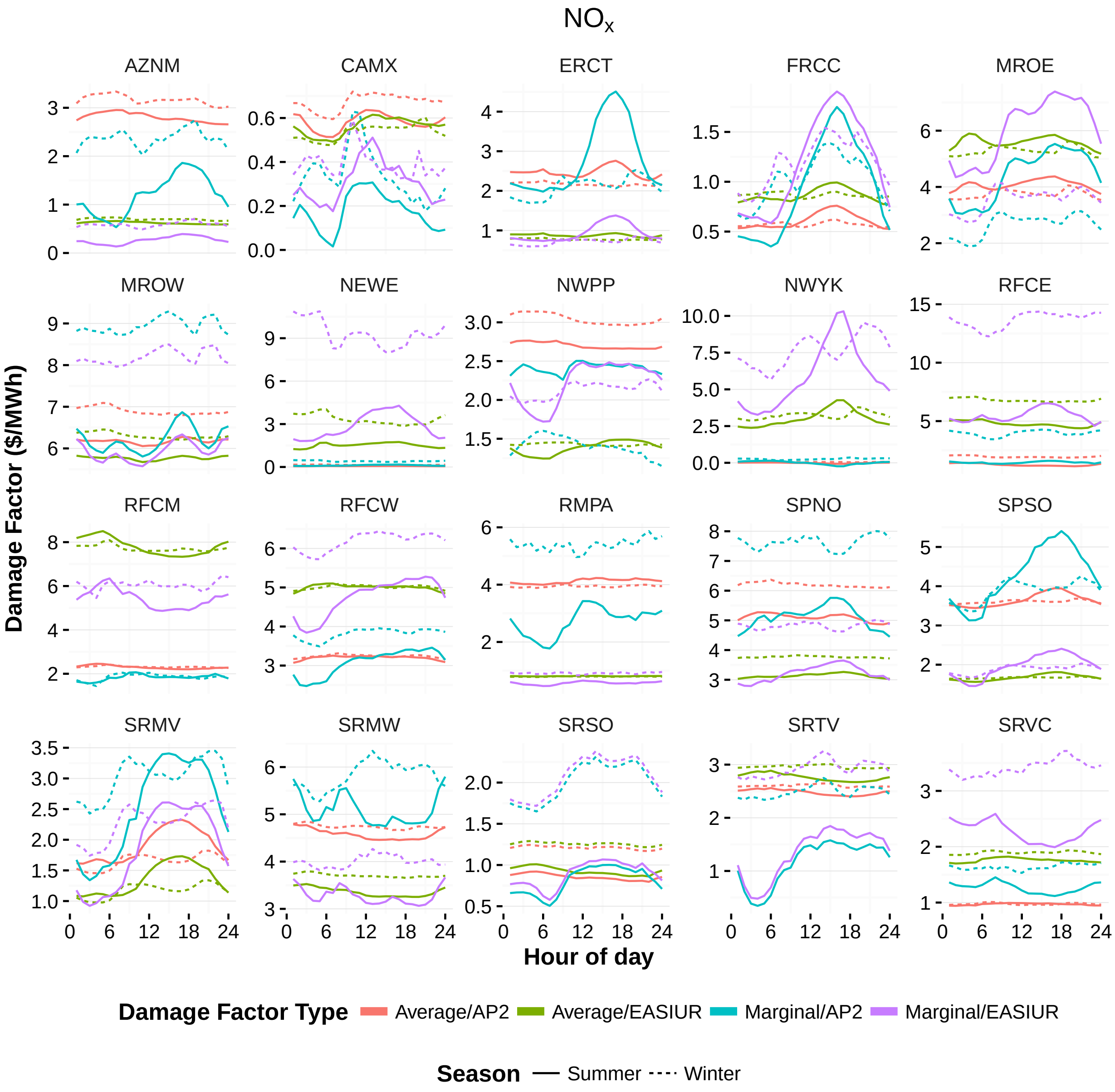


Figure A.15: NO_x seasonal hourly marginal and average damage factors by subregion. Note that the **y-axis scale varies and does not start at zero** to aid visualization of the patterns among the various damage factor estimates within each subregion. See Figure A.36 for an accurate inter-regional comparison, and Figures A.17 – A.35 for a full-scale representation of the damage estimates within each region.

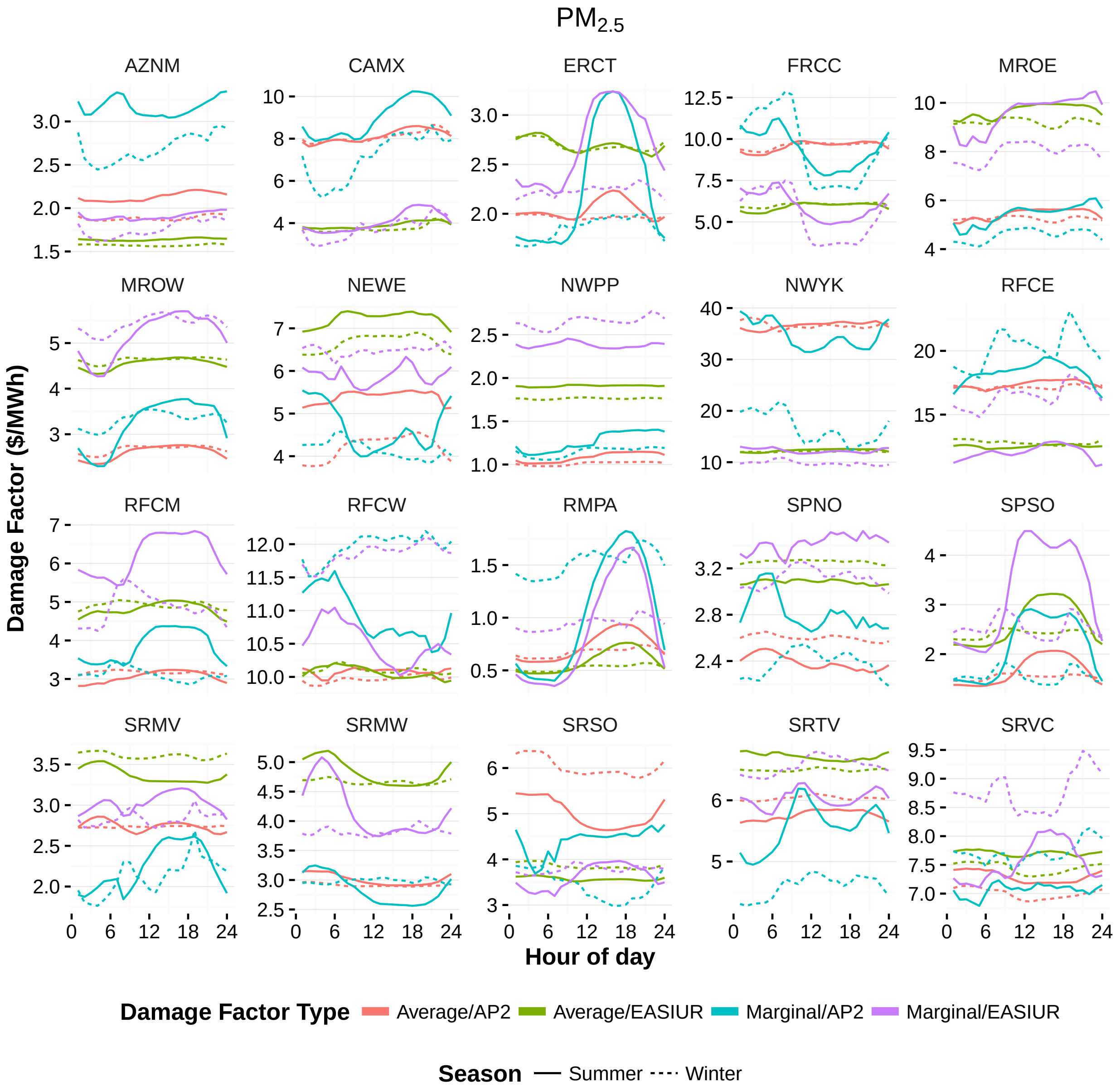


Figure A.16: PM$_{2.5}$ seasonal hourly marginal and average damage factors by subregion. Note that the **y-axis scale varies and does not start at zero** to aid visualization of the patterns among the various damage factor estimates within each subregion. See Figure A.36 for an accurate inter-regional comparison, and Figures A.17 – A.35 for a full-scale representation of the damage estimates within each region.

Our final set of figures makes the inter-regional and seasonal comparisons clear and shows the relative contribution of each pollutant to total damages. Figures A.17 – A.35 show EASIUR and AP2 marginal and average damage factors (designated in the panel titles as *MDF* and *ADF*, respectively) by hour and season for each subregion. Figure A.36 shows EASIUR marginal damage factors for all subregions with a common fixed y-axes to provide a sense of the differences in scale among the subregions.

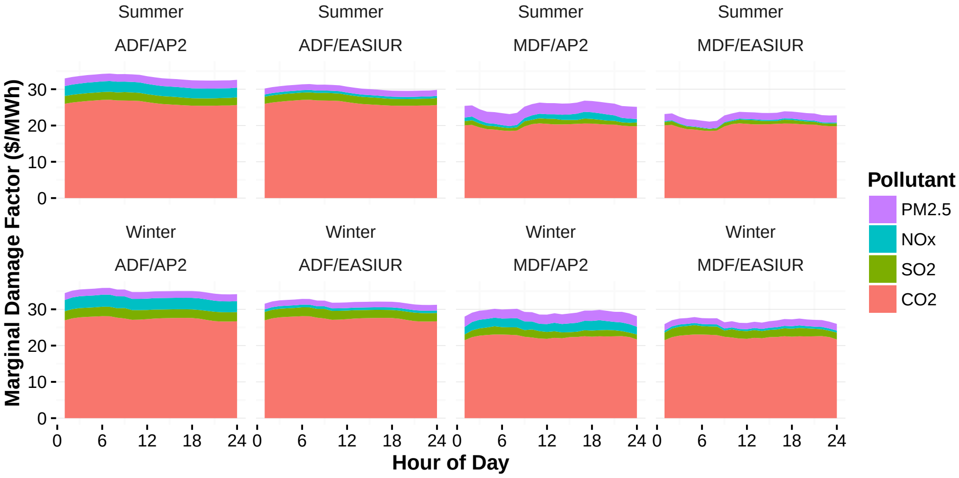


Figure A.17: Seasonal hourly average and marginal damage factors, AZNM subregion.

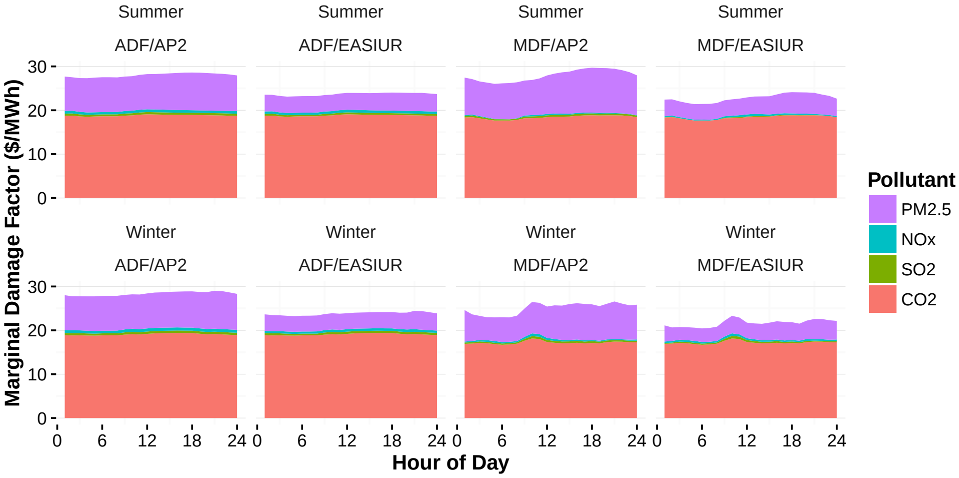


Figure A.18: Seasonal hourly average and marginal damage factors, CAMX subregion.

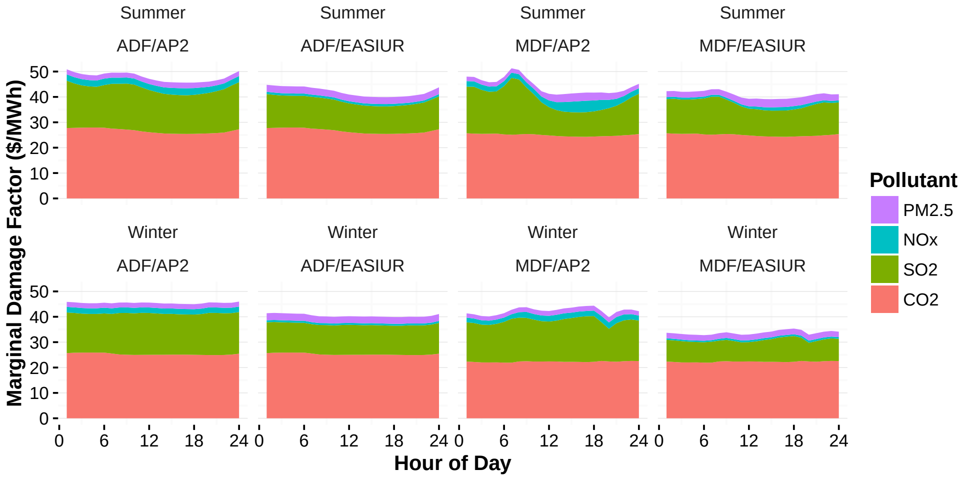


Figure A.19: Seasonal hourly average and marginal damage factors, ERCT subregion.

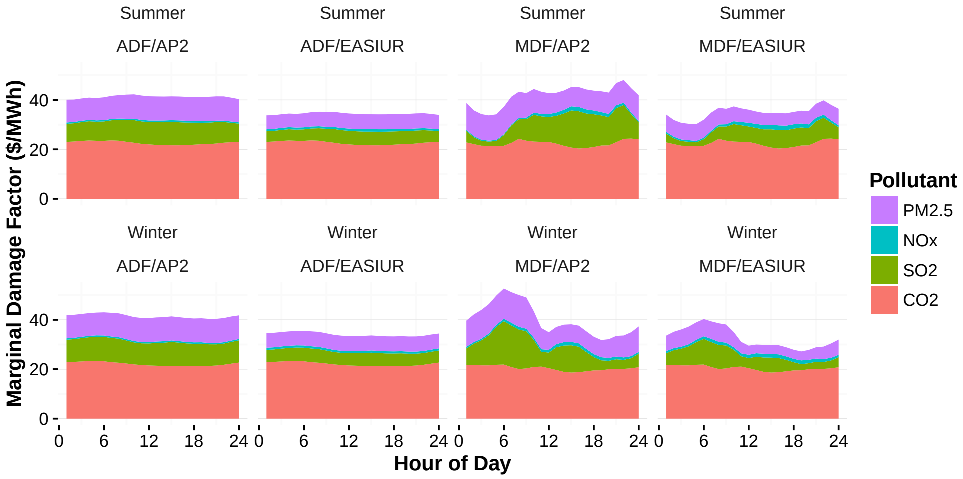


Figure A.20: Seasonal hourly average and marginal damage factors, FRCC subregion.

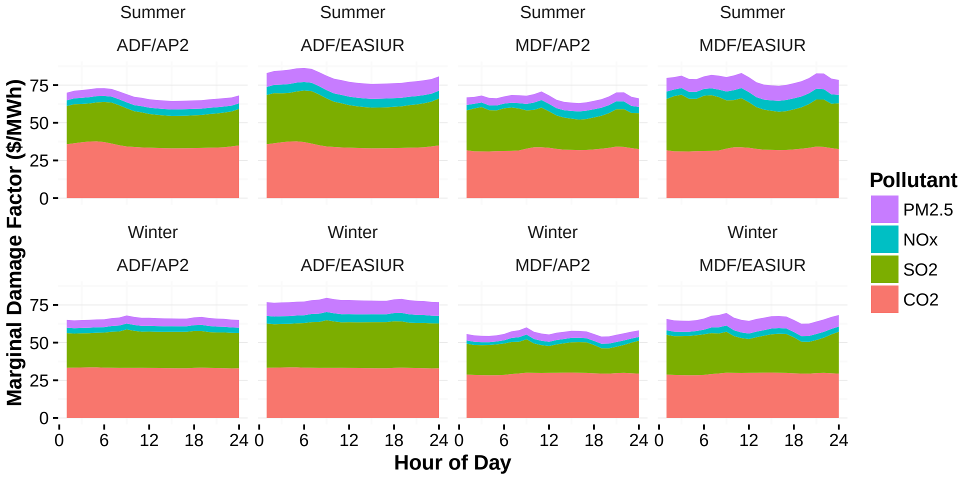


Figure A.21: Seasonal hourly average and marginal damage factors, MROE subregion.

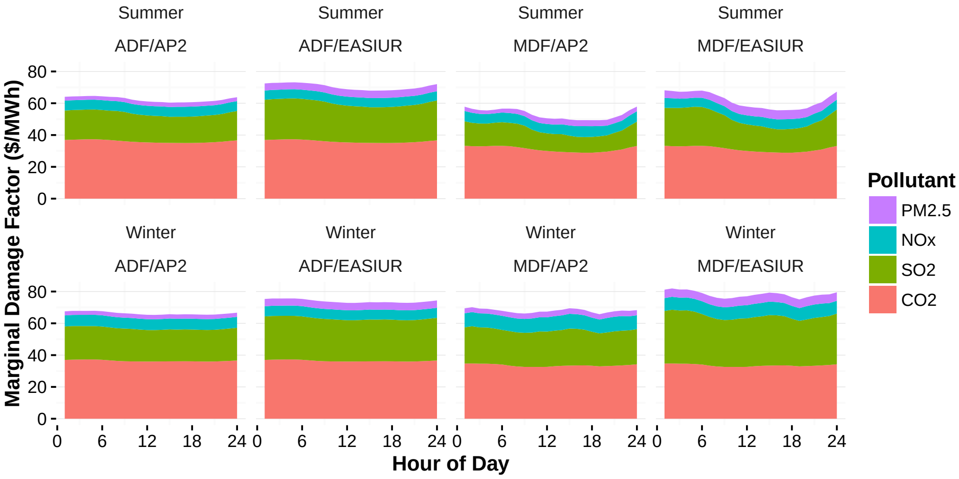


Figure A.22: Seasonal hourly average and marginal damage factors, MROW subregion.

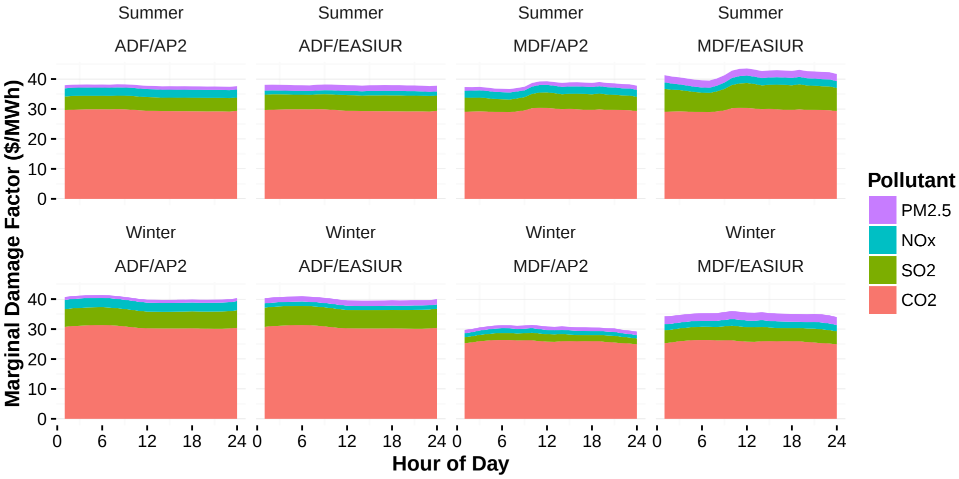


Figure A.23: Seasonal hourly average and marginal damage factors, NWPP subregion.

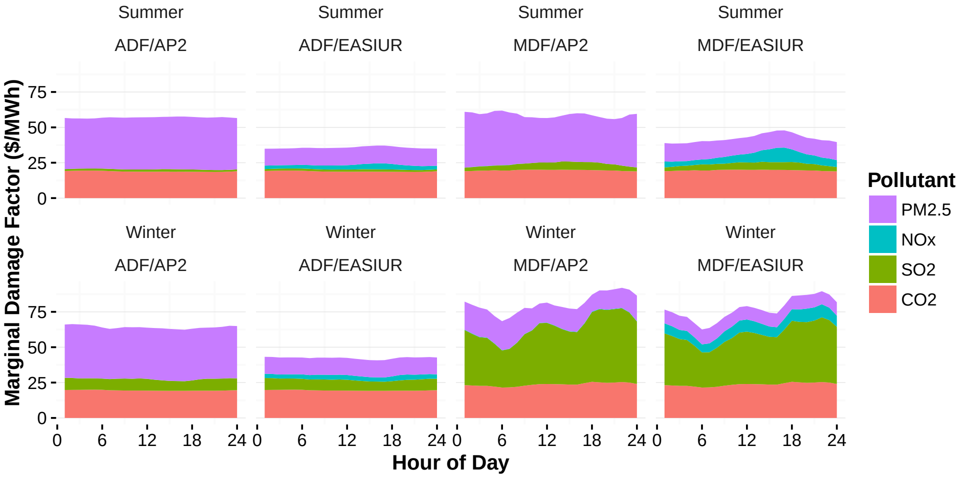


Figure A.24: Seasonal hourly average and marginal damage factors, NWYK subregion.

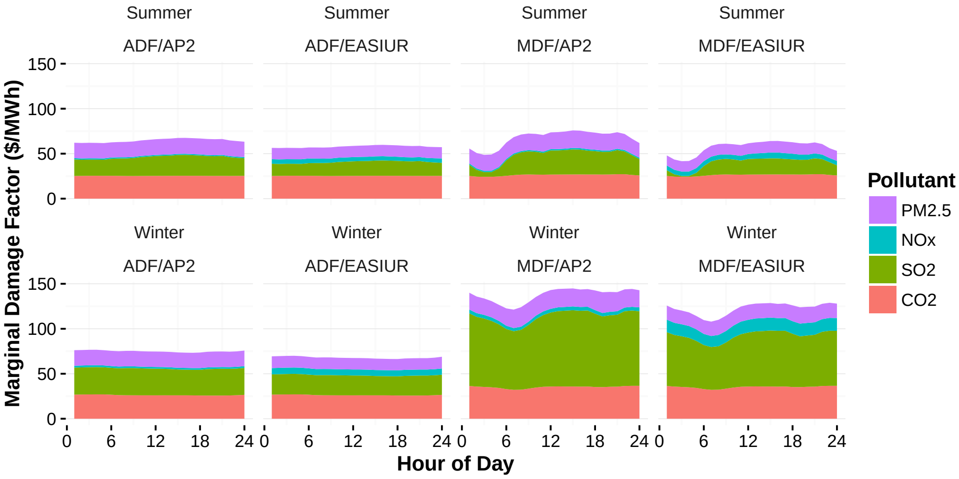


Figure A.25: Seasonal hourly average and marginal damage factors, RFCE subregion.

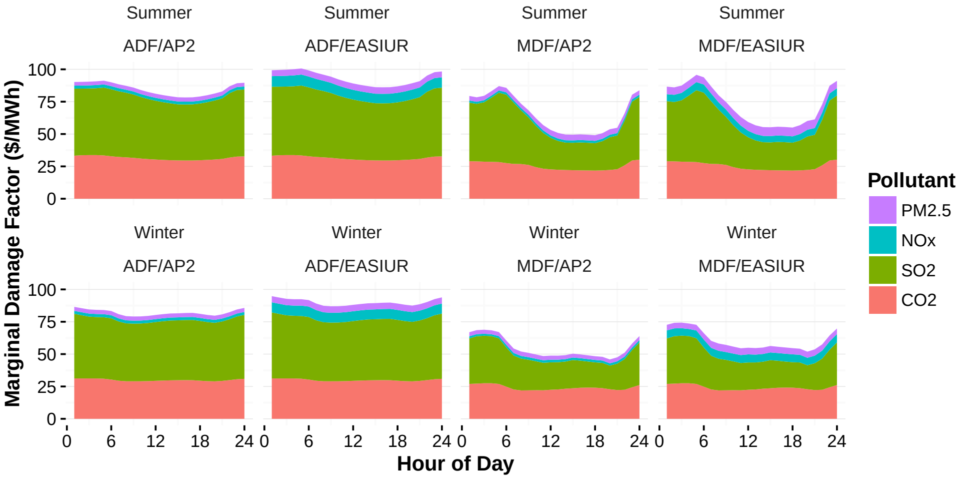


Figure A.26: Seasonal hourly average and marginal damage factors, RFCM subregion.

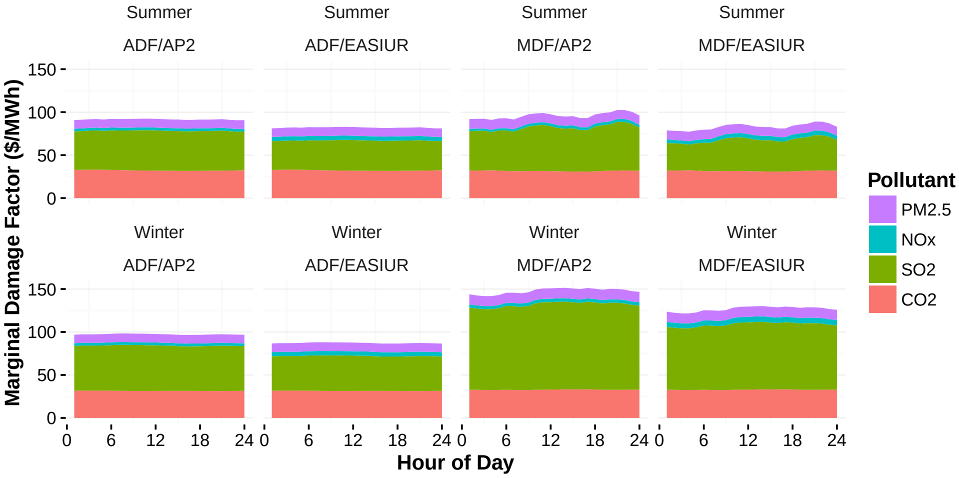


Figure A.27: Seasonal hourly average and marginal damage factors, RFCW subregion.

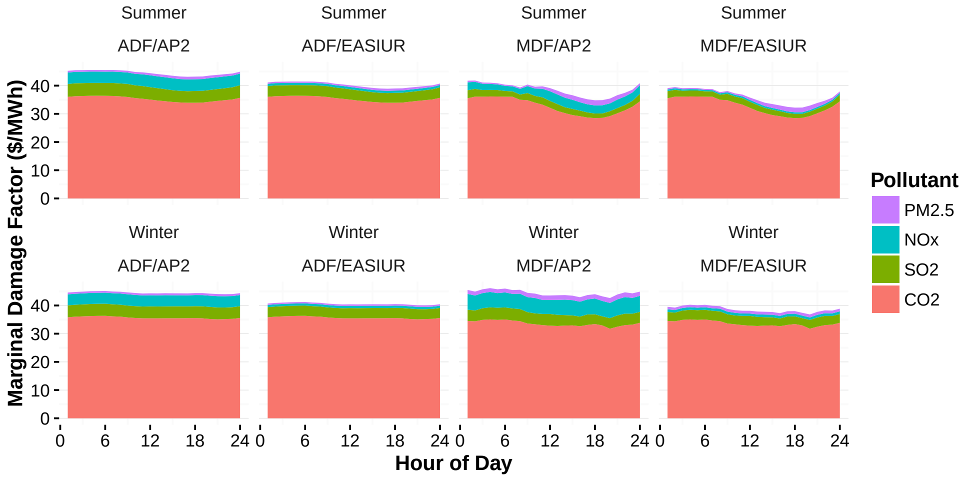


Figure A.28: Seasonal hourly average and marginal damage factors, RMPA subregion.

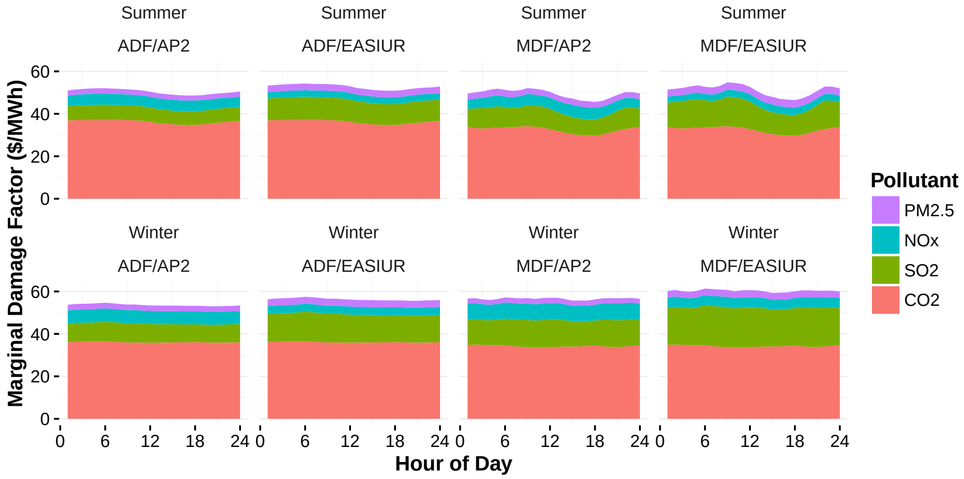


Figure A.29: Seasonal hourly average and marginal damage factors, SPNO subregion.

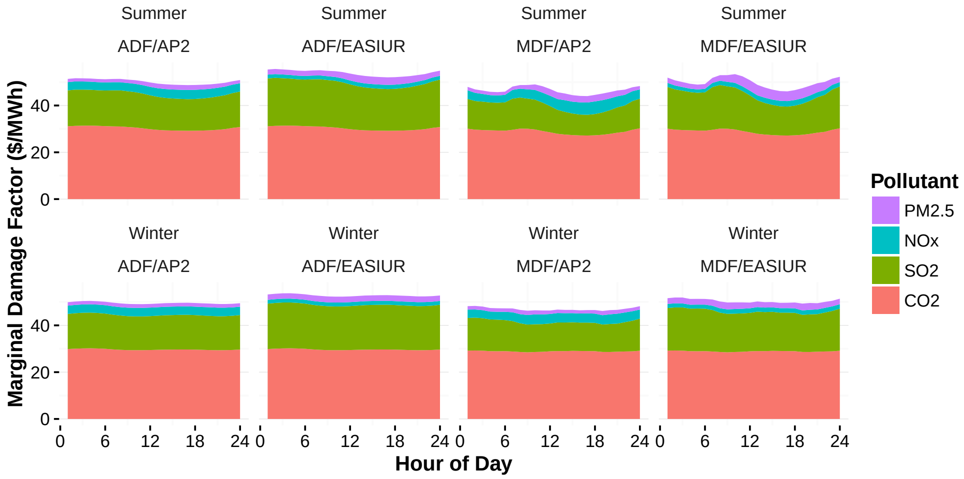


Figure A.30: Seasonal hourly average and marginal damage factors, SPSO subregion.

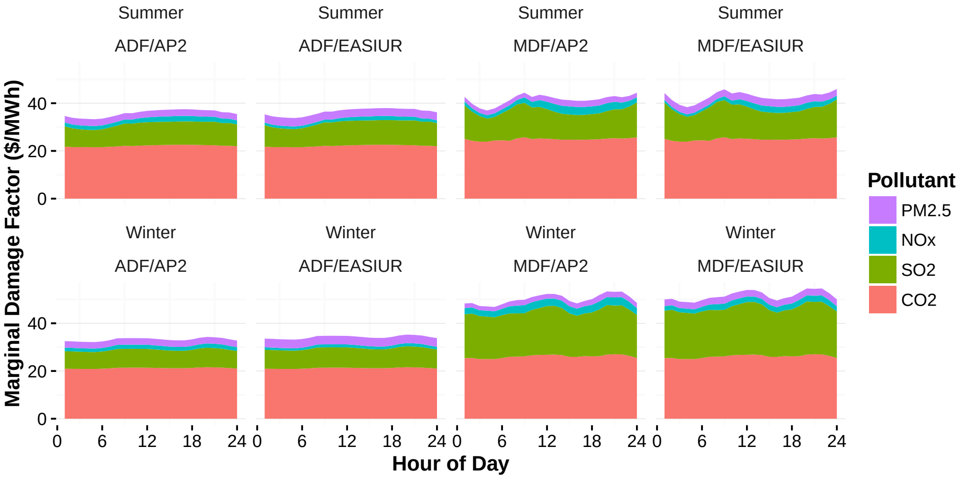


Figure A.31: Seasonal hourly average and marginal damage factors, SRMV subregion.

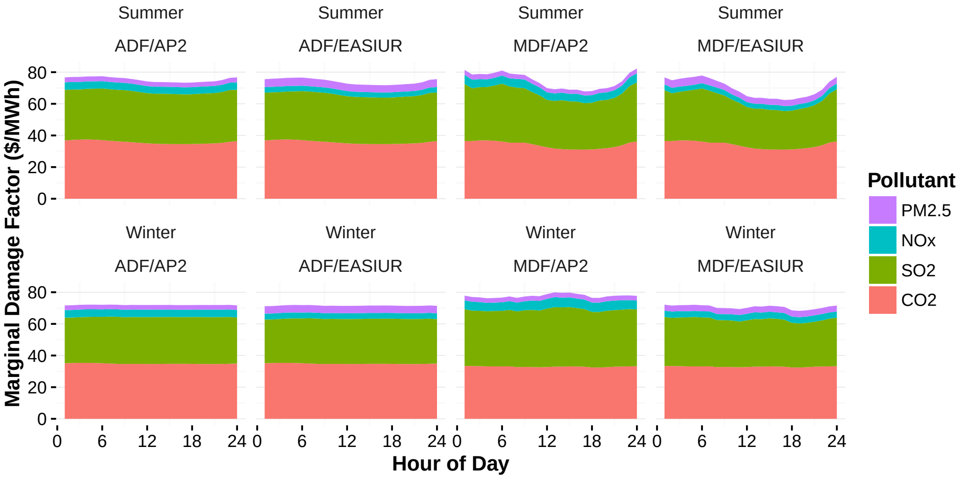


Figure A.32: Seasonal hourly average and marginal damage factors, SRMW subregion.

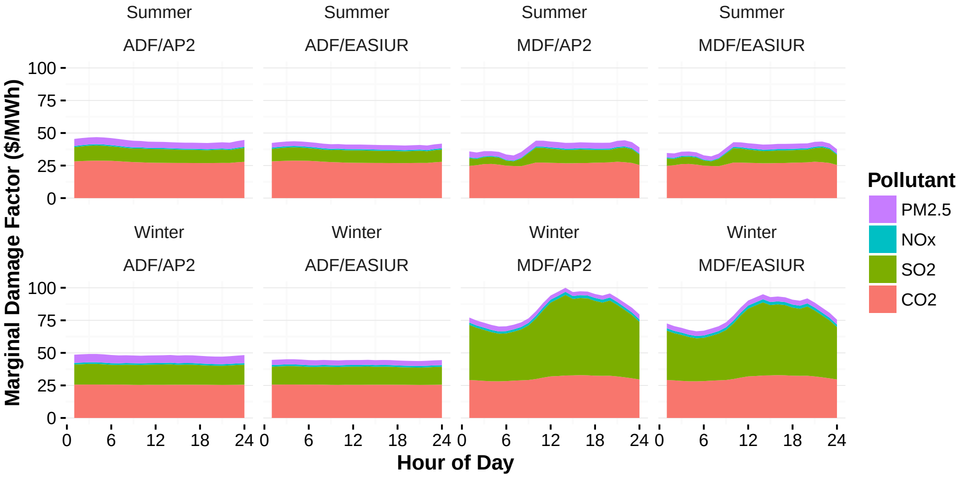


Figure A.33: Seasonal hourly average and marginal damage factors, SRSO subregion.

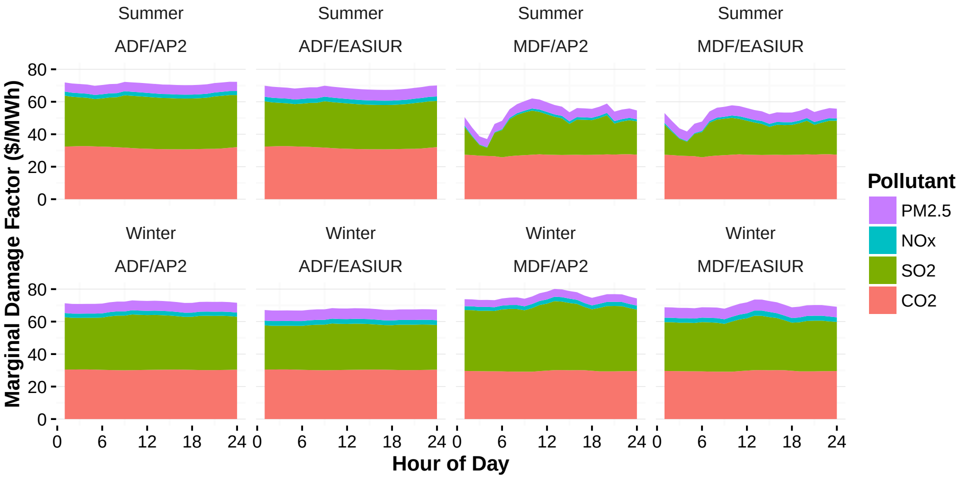


Figure A.34: Seasonal hourly average and marginal damage factors, SRTV subregion.

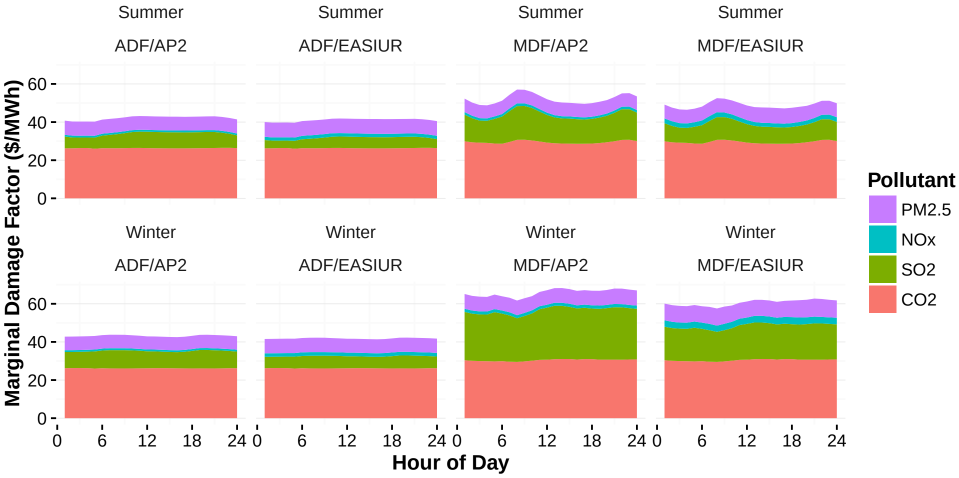


Figure A.35: Seasonal hourly average and marginal damage factors, SRVC subregion.

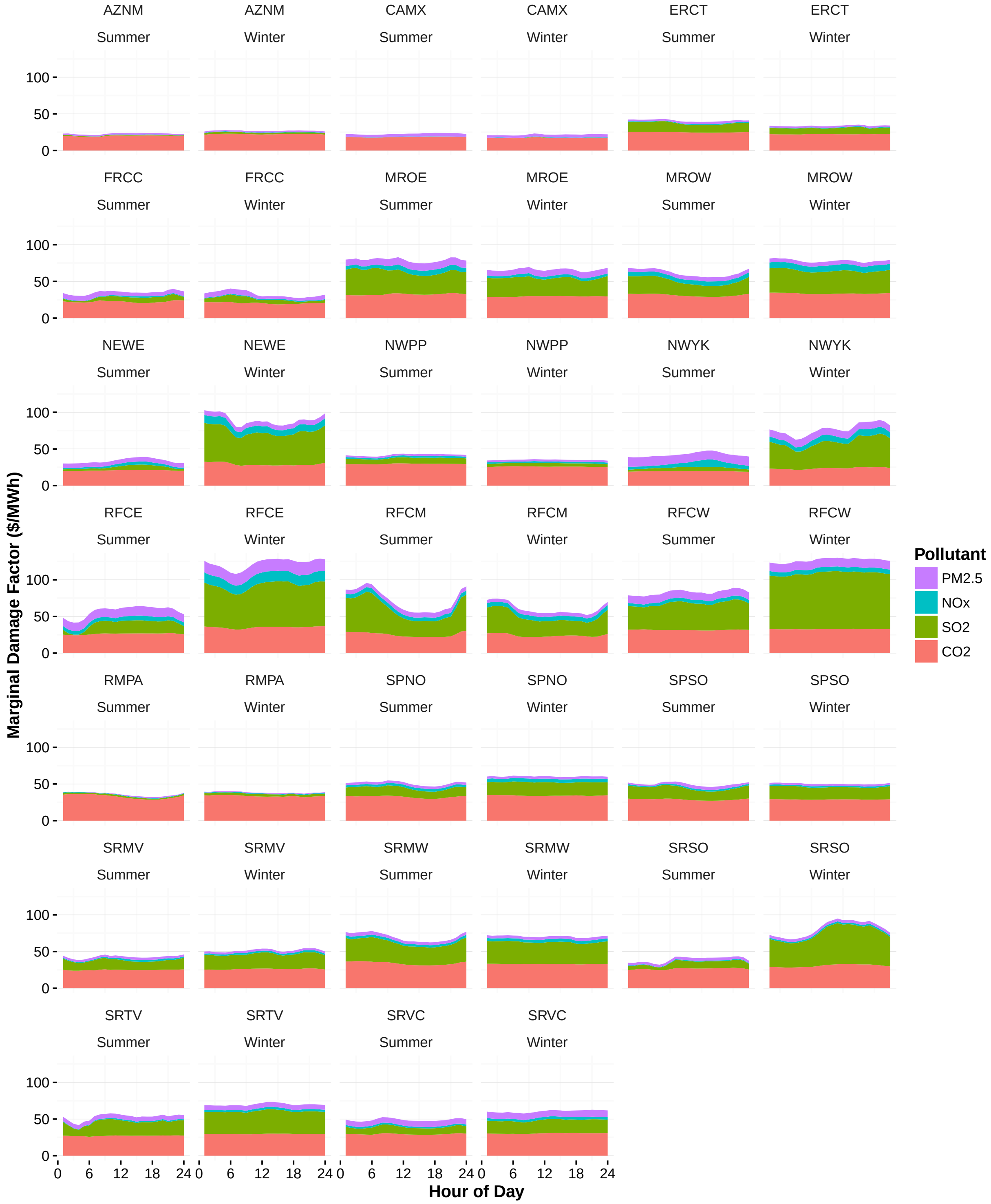


Figure A.36: EASIUR hourly marginal damage factors by subregion and season. Y-axis is fixed to enable comparison among regions.

Appendix B

Supporting Information

This appendix contains supporting tables referenced in the body of the dissertation.

Table B.1: Correlation between MDFs and LMPs.

MDF	*LMP*															
	AL	**AR**	**AZ**	**CA**	**CO**	**CT**	**DE**	**FL**	**GA**	**IA**	**ID**	**IL**	**IN**	**KS**	**KY**	**LA**
AL	0.10	0.07	-0.03	-0.02	-0.03	0.07	0.07	0.09	0.09	0.08	-0.01	0.07	0.04	0.09	0.05	0.10
AR	0.07	0.05	0.04	0.04	0.03	0.02	0.04	0.08	0.08	0.09	0.05	0.09	0.09	0.08	0.08	0.04
AZ	0.04	0.03	0.01	0.01	0.01	0.02	0.03	0.04	0.04	0.02	0.02	0.03	0.03	0.02	0.03	0.05
CA	0.13	0.14	0.09	0.13	0.11	0.08	0.05	0.14	0.14	0.19	0.10	0.15	0.11	0.17	0.12	0.04
CO	-0.17	-0.18	-0.06	-0.10	-0.08	-0.10	-0.07	-0.18	-0.18	-0.22	-0.07	-0.18	-0.16	-0.21	-0.17	-0.06
CT	0.00	0.00	-0.03	-0.04	-0.03	0.02	0.02	0.00	0.00	-0.04	-0.02	-0.02	-0.04	-0.02	-0.03	0.04
DE	0.04	0.02	-0.01	-0.02	-0.02	0.03	0.04	0.04	0.04	0.01	-0.01	0.02	0.01	0.02	0.01	0.06
FL	-0.02	0.01	0.03	0.03	0.03	-0.02	-0.01	-0.01	-0.01	0.01	0.02	0.01	0.01	0.01	0.01	-0.05
GA	0.10	0.07	-0.03	-0.02	-0.03	0.07	0.07	0.09	0.09	0.08	-0.01	0.07	0.04	0.09	0.05	0.10
IA	-0.07	-0.08	-0.03	-0.06	-0.05	-0.03	-0.02	-0.08	-0.08	-0.12	-0.04	-0.10	-0.10	-0.11	-0.10	0.01
ID	0.06	0.07	0.03	0.06	0.05	0.02	0.01	0.07	0.07	0.12	0.04	0.10	0.08	0.11	0.08	-0.01
IL	-0.17	-0.18	-0.01	-0.05	-0.03	-0.11	-0.08	-0.17	-0.17	-0.19	-0.03	-0.17	-0.14	-0.20	-0.14	-0.08
IN	0.03	0.01	-0.03	-0.04	-0.04	0.03	0.04	0.02	0.02	0.00	-0.02	0.01	0.00	0.01	0.01	0.05
KS	-0.03	-0.05	-0.04	-0.06	-0.05	-0.01	0.01	-0.04	-0.04	-0.06	-0.04	-0.04	-0.05	-0.05	-0.05	0.02
KY	0.06	0.04	-0.02	-0.02	-0.02	0.04	0.04	0.06	0.06	0.04	0.00	0.04	0.03	0.05	0.02	0.06
LA	0.07	0.05	0.04	0.04	0.03	0.02	0.04	0.08	0.08	0.09	0.05	0.09	0.09	0.08	0.08	0.04
MA	0.00	0.00	-0.03	-0.04	-0.03	0.02	0.02	0.00	0.00	-0.04	-0.02	-0.02	-0.04	-0.02	-0.03	0.04
MD	0.04	0.02	-0.01	-0.02	-0.02	0.03	0.04	0.04	0.04	0.01	-0.01	0.02	0.01	0.02	0.01	0.06
ME	0.00	0.00	-0.03	-0.04	-0.03	0.02	0.02	0.00	0.00	-0.04	-0.02	-0.02	-0.04	-0.02	-0.03	0.04
MI	-0.25	-0.24	0.01	-0.03	0.00	-0.16	-0.14	-0.25	-0.25	-0.28	-0.02	-0.26	-0.21	-0.28	-0.21	-0.14
MN	-0.07	-0.08	-0.03	-0.06	-0.05	-0.03	-0.02	-0.08	-0.08	-0.12	-0.04	-0.10	-0.10	-0.11	-0.10	0.01
MO	-0.17	-0.18	-0.01	-0.05	-0.03	-0.11	-0.08	-0.17	-0.17	-0.19	-0.03	-0.17	-0.14	-0.20	-0.14	-0.08
MS	0.07	0.05	0.04	0.04	0.03	0.02	0.04	0.08	0.08	0.09	0.05	0.09	0.09	0.08	0.08	0.04
MT	0.06	0.07	0.03	0.06	0.05	0.02	0.01	0.07	0.07	0.12	0.04	0.10	0.08	0.11	0.08	-0.01
NC	0.04	0.02	-0.01	-0.02	-0.02	0.04	0.04	0.03	0.03	0.01	-0.01	0.02	0.00	0.03	0.00	0.06
ND	-0.07	-0.08	-0.03	-0.06	-0.05	-0.03	-0.02	-0.08	-0.08	-0.12	-0.04	-0.10	-0.10	-0.11	-0.10	0.01
NE	-0.07	-0.08	-0.03	-0.06	-0.05	-0.03	-0.02	-0.08	-0.08	-0.12	-0.04	-0.10	-0.10	-0.11	-0.10	0.01
NH	0.00	0.00	-0.03	-0.04	-0.03	0.02	0.02	0.00	0.00	-0.04	-0.02	-0.02	-0.04	-0.02	-0.03	0.04
NJ	0.04	0.02	-0.01	-0.02	-0.02	0.03	0.04	0.04	0.04	0.01	-0.01	0.02	0.01	0.02	0.01	0.06
NM	0.04	0.03	0.01	0.01	0.01	0.02	0.03	0.04	0.04	0.02	0.02	0.03	0.03	0.02	0.03	0.05
NV	0.06	0.07	0.03	0.06	0.05	0.02	0.01	0.07	0.07	0.12	0.04	0.10	0.08	0.11	0.08	-0.01
NY	0.06	0.04	-0.02	-0.02	-0.02	0.05	0.04	0.05	0.05	0.02	-0.01	0.03	0.00	0.04	0.01	0.07
OH	0.03	0.01	-0.03	-0.04	-0.04	0.03	0.04	0.02	0.02	0.00	-0.02	0.01	0.00	0.01	0.01	0.05
OK	0.00	-0.01	0.00	-0.01	-0.01	0.01	0.00	0.01	0.01	0.02	-0.01	0.03	-0.01	0.03	0.00	0.00
OR	0.06	0.07	0.03	0.06	0.05	0.02	0.01	0.07	0.07	0.12	0.04	0.10	0.08	0.11	0.08	-0.01
PA	0.04	0.02	-0.01	-0.02	-0.02	0.03	0.04	0.04	0.04	0.01	-0.01	0.02	0.01	0.02	0.01	0.06
RI	0.00	0.00	-0.03	-0.04	-0.03	0.02	0.02	0.00	0.00	-0.04	-0.02	-0.02	-0.04	-0.02	-0.03	0.04
SC	0.04	0.02	-0.01	-0.02	-0.02	0.04	0.04	0.03	0.03	0.01	-0.01	0.02	0.00	0.03	0.00	0.06
SD	-0.07	-0.08	-0.03	-0.06	-0.05	-0.03	-0.02	-0.08	-0.08	-0.12	-0.04	-0.10	-0.10	-0.11	-0.10	0.01
TN	0.06	0.04	-0.02	-0.02	-0.02	0.04	0.04	0.06	0.06	0.04	0.00	0.04	0.03	0.05	0.02	0.06
TX	-0.03	-0.03	0.05	0.05	0.05	-0.04	-0.03	-0.02	-0.02	0.00	0.04	-0.01	0.02	-0.02	0.01	-0.05
UT	0.06	0.07	0.03	0.06	0.05	0.02	0.01	0.07	0.07	0.12	0.04	0.10	0.08	0.11	0.08	-0.01
VA	0.04	0.02	-0.01	-0.02	-0.02	0.04	0.04	0.03	0.03	0.01	-0.01	0.02	0.00	0.03	0.00	0.06
VT	0.00	0.00	-0.03	-0.04	-0.03	0.02	0.02	0.00	0.00	-0.04	-0.02	-0.02	-0.04	-0.02	-0.03	0.04
WA	0.06	0.07	0.03	0.06	0.05	0.02	0.01	0.07	0.07	0.12	0.04	0.10	0.08	0.11	0.08	-0.01
WI	-0.03	-0.03	0.05	0.05	0.05	-0.04	-0.03	-0.01	-0.01	0.01	0.04	0.01	0.04	-0.02	0.03	-0.05
WV	0.03	0.01	-0.03	-0.04	-0.04	0.03	0.04	0.02	0.02	0.00	-0.02	0.01	0.00	0.01	0.01	0.05
WY	0.06	0.07	0.03	0.06	0.05	0.02	0.01	0.07	0.07	0.12	0.04	0.10	0.08	0.11	0.08	-0.01

Table B.2: Correlation between MDFs and LMPs, continued.

MDF	*LMP*															
	MA	**MD**	**ME**	**MI**	**MN**	**MO**	**MS**	**MT**	**NC**	**ND**	**NE**	**NH**	**NJ**	**NM**	**NV**	**NY**
AL	0.07	0.07	0.08	0.05	0.10	0.10	0.09	-0.01	0.06	0.07	0.09	0.08	0.04	-0.03	-0.03	0.05
AR	0.02	0.06	0.02	0.10	0.12	0.09	0.04	0.04	0.10	0.13	0.08	0.02	0.03	0.03	0.04	0.03
AZ	0.02	0.04	0.02	0.03	0.05	0.04	0.04	0.02	0.04	0.05	0.02	0.02	0.01	0.01	0.01	0.01
CA	0.08	0.06	0.08	0.12	0.19	0.16	0.07	0.11	0.13	0.19	0.17	0.08	0.07	0.10	0.11	0.11
CO	-0.10	-0.11	-0.09	-0.17	-0.22	-0.20	-0.09	-0.09	-0.18	-0.22	-0.21	-0.09	-0.10	-0.06	-0.08	-0.13
CT	0.02	0.00	0.02	-0.03	-0.01	0.00	0.03	-0.03	-0.03	-0.03	-0.02	0.02	-0.01	-0.03	-0.03	-0.01
DE	0.03	0.03	0.03	0.02	0.04	0.04	0.05	-0.01	0.02	0.02	0.02	0.03	0.01	-0.02	-0.02	0.01
FL	-0.01	-0.01	-0.01	0.00	0.01	0.00	-0.05	0.03	0.01	0.01	0.01	-0.01	0.00	0.03	0.03	-0.01
GA	0.07	0.07	0.08	0.05	0.10	0.10	0.09	-0.01	0.06	0.07	0.09	0.08	0.04	-0.03	-0.03	0.05
IA	-0.03	-0.05	-0.03	-0.10	-0.11	-0.09	-0.01	-0.05	-0.10	-0.12	-0.11	-0.03	-0.05	-0.04	-0.04	-0.07
ID	0.03	0.03	0.02	0.08	0.10	0.09	0.01	0.05	0.09	0.11	0.11	0.02	0.04	0.04	0.04	0.06
IL	-0.11	-0.12	-0.11	-0.14	-0.20	-0.19	-0.10	-0.04	-0.16	-0.17	-0.20	-0.11	-0.10	-0.02	-0.02	-0.13
IN	0.03	0.04	0.03	0.01	0.03	0.03	0.05	-0.03	0.01	0.00	0.01	0.03	0.01	-0.04	-0.04	0.00
KS	-0.01	-0.01	0.00	-0.04	-0.04	-0.04	0.00	-0.05	-0.04	-0.06	-0.05	0.00	-0.02	-0.05	-0.05	-0.04
KY	0.04	0.05	0.05	0.04	0.07	0.07	0.06	-0.01	0.04	0.04	0.05	0.05	0.02	-0.02	-0.02	0.02
LA	0.02	0.06	0.02	0.10	0.12	0.09	0.04	0.04	0.10	0.13	0.08	0.02	0.03	0.03	0.04	0.03
MA	0.02	0.00	0.02	-0.03	-0.01	0.00	0.03	-0.03	-0.03	-0.03	-0.02	0.02	-0.01	-0.03	-0.03	-0.01
MD	0.03	0.03	0.03	0.02	0.04	0.04	0.05	-0.01	0.02	0.02	0.02	0.03	0.01	-0.02	-0.02	0.01
ME	0.02	0.00	0.02	-0.03	-0.01	0.00	0.03	-0.03	-0.03	-0.03	-0.02	0.02	-0.01	-0.03	-0.03	-0.01
MI	-0.16	-0.19	-0.16	-0.22	-0.30	-0.28	-0.17	-0.03	-0.24	-0.27	-0.28	-0.16	-0.14	0.01	0.00	-0.17
MN	-0.03	-0.05	-0.03	-0.10	-0.11	-0.09	-0.01	-0.05	-0.10	-0.12	-0.11	-0.03	-0.05	-0.04	-0.04	-0.07
MO	-0.11	-0.12	-0.11	-0.14	-0.20	-0.19	-0.10	-0.04	-0.16	-0.17	-0.20	-0.11	-0.10	-0.02	-0.02	-0.13
MS	0.02	0.06	0.02	0.10	0.12	0.09	0.04	0.04	0.10	0.13	0.08	0.02	0.03	0.03	0.04	0.03
MT	0.03	0.03	0.02	0.08	0.10	0.09	0.01	0.05	0.09	0.11	0.11	0.02	0.04	0.04	0.04	0.06
NC	0.03	0.03	0.04	0.01	0.04	0.04	0.05	-0.01	0.01	0.01	0.03	0.04	0.01	-0.02	-0.02	0.01
ND	-0.03	-0.05	-0.03	-0.10	-0.11	-0.09	-0.01	-0.05	-0.10	-0.12	-0.11	-0.03	-0.05	-0.04	-0.04	-0.07
NE	-0.03	-0.05	-0.03	-0.10	-0.11	-0.09	-0.01	-0.05	-0.10	-0.12	-0.11	-0.03	-0.05	-0.04	-0.04	-0.07
NH	0.02	0.00	0.02	-0.03	-0.01	0.00	0.03	-0.03	-0.03	-0.03	-0.02	0.02	-0.01	-0.03	-0.03	-0.01
NJ	0.03	0.03	0.03	0.02	0.04	0.04	0.05	-0.01	0.02	0.02	0.02	0.03	0.01	-0.02	-0.02	0.01
NM	0.02	0.04	0.02	0.03	0.05	0.04	0.04	0.02	0.04	0.05	0.02	0.02	0.01	0.01	0.01	0.01
NV	0.03	0.03	0.02	0.08	0.10	0.09	0.01	0.05	0.09	0.11	0.11	0.02	0.04	0.04	0.04	0.06
NY	0.05	0.03	0.05	0.01	0.05	0.05	0.07	-0.02	0.02	0.02	0.04	0.05	0.02	-0.02	-0.02	0.02
OH	0.03	0.04	0.03	0.01	0.03	0.03	0.05	-0.03	0.01	0.00	0.01	0.03	0.01	-0.04	-0.04	0.00
OK	0.01	-0.01	0.02	0.00	0.02	0.02	0.00	-0.01	0.00	0.03	0.03	0.02	0.00	0.00	0.00	0.00
OR	0.03	0.03	0.02	0.08	0.10	0.09	0.01	0.05	0.09	0.11	0.11	0.02	0.04	0.04	0.04	0.06
PA	0.03	0.03	0.03	0.02	0.04	0.04	0.05	-0.01	0.02	0.02	0.02	0.03	0.01	-0.02	-0.02	0.01
RI	0.02	0.00	0.02	-0.03	-0.01	0.00	0.03	-0.03	-0.03	-0.03	-0.02	0.02	-0.01	-0.03	-0.03	-0.01
SC	0.03	0.03	0.04	0.01	0.04	0.04	0.05	-0.01	0.01	0.01	0.03	0.04	0.01	-0.02	-0.02	0.01
SD	-0.03	-0.05	-0.03	-0.10	-0.11	-0.09	-0.01	-0.05	-0.10	-0.12	-0.11	-0.03	-0.05	-0.04	-0.04	-0.07
TN	0.04	0.05	0.05	0.04	0.07	0.07	0.06	-0.01	0.04	0.04	0.05	0.05	0.02	-0.02	-0.02	0.02
TX	-0.04	-0.02	-0.04	0.01	-0.01	-0.02	-0.05	0.04	0.01	0.01	-0.02	-0.04	-0.02	0.05	0.05	-0.01
UT	0.03	0.03	0.02	0.08	0.10	0.09	0.01	0.05	0.09	0.11	0.11	0.02	0.04	0.04	0.04	0.06
VA	0.03	0.03	0.04	0.01	0.04	0.04	0.05	-0.01	0.01	0.01	0.03	0.04	0.01	-0.02	-0.02	0.01
VT	0.02	0.00	0.02	-0.03	-0.01	0.00	0.03	-0.03	-0.03	-0.03	-0.02	0.02	-0.01	-0.03	-0.03	-0.01
WA	0.03	0.03	0.02	0.08	0.10	0.09	0.01	0.05	0.09	0.11	0.11	0.02	0.04	0.04	0.04	0.06
WI	-0.04	-0.01	-0.04	0.02	0.00	-0.02	-0.04	0.05	0.02	0.03	-0.02	-0.04	-0.01	0.05	0.05	-0.01
WV	0.03	0.04	0.03	0.01	0.03	0.03	0.05	-0.03	0.01	0.00	0.01	0.03	0.01	-0.04	-0.04	0.00
WY	0.03	0.03	0.02	0.08	0.10	0.09	0.01	0.05	0.09	0.11	0.11	0.02	0.04	0.04	0.04	0.06

Table B.3: Correlation between MDFs and LMPs, continued.

MDF	*LMP*															
	OH	**OK**	**OR**	**PA**	**RI**	**SC**	**SD**	**TN**	**TX**	**UT**	**VA**	**VT**	**WA**	**WI**	**WV**	**WY**
AL	0.05	0.09	0.01	0.05	0.07	0.09	0.08	0.08	0.07	0.03	0.04	0.08	-0.01	0.07	0.04	-0.01
AR	0.08	0.08	0.01	0.04	0.02	0.08	0.09	0.09	0.05	0.04	0.05	0.02	0.05	0.11	0.06	0.04
AZ	0.03	0.02	0.01	0.02	0.02	0.04	0.02	0.03	0.05	0.02	0.02	0.02	0.02	0.03	0.02	0.02
CA	0.11	0.17	0.06	0.07	0.08	0.14	0.19	0.15	0.10	0.08	0.08	0.08	0.11	0.16	0.09	0.08
CO	-0.16	-0.21	-0.06	-0.09	-0.09	-0.18	-0.22	-0.20	-0.11	-0.06	-0.13	-0.09	-0.08	-0.20	-0.14	-0.06
CT	-0.02	-0.02	0.00	0.00	0.02	0.00	-0.03	-0.02	0.01	0.01	-0.02	0.02	-0.03	-0.03	-0.02	-0.02
DE	0.01	0.02	0.00	0.02	0.03	0.04	0.01	0.02	0.04	0.02	0.01	0.03	-0.01	0.02	0.01	0.00
FL	0.00	0.01	-0.01	0.00	-0.02	-0.01	0.02	0.02	-0.02	-0.04	0.00	-0.01	0.03	0.01	0.00	0.02
GA	0.05	0.09	0.01	0.05	0.07	0.09	0.08	0.08	0.07	0.03	0.04	0.08	-0.01	0.07	0.04	-0.01
IA	-0.09	-0.11	-0.03	-0.04	-0.03	-0.08	-0.12	-0.10	-0.04	-0.01	-0.08	-0.03	-0.05	-0.12	-0.08	-0.03
ID	0.07	0.11	0.03	0.03	0.02	0.07	0.12	0.10	0.03	0.02	0.06	0.02	0.05	0.11	0.06	0.03
IL	-0.15	-0.20	-0.06	-0.10	-0.11	-0.17	-0.19	-0.18	-0.10	-0.04	-0.12	-0.11	-0.04	-0.18	-0.13	-0.02
IN	0.01	0.01	0.00	0.02	0.03	0.02	0.00	0.01	0.02	0.00	0.01	0.03	-0.03	0.01	0.01	-0.02
KS	-0.04	-0.05	-0.03	-0.01	-0.01	-0.04	-0.06	-0.05	-0.03	-0.02	-0.04	-0.01	-0.05	-0.05	-0.04	-0.03
KY	0.03	0.05	0.01	0.03	0.04	0.06	0.04	0.05	0.05	0.02	0.02	0.04	-0.01	0.05	0.02	0.00
LA	0.08	0.08	0.01	0.04	0.02	0.08	0.09	0.09	0.05	0.04	0.05	0.02	0.05	0.11	0.06	0.04
MA	-0.02	-0.02	0.00	0.00	0.02	0.00	-0.03	-0.02	0.01	0.01	-0.02	0.02	-0.03	-0.03	-0.02	-0.02
MD	0.01	0.02	0.00	0.02	0.03	0.04	0.01	0.02	0.04	0.02	0.01	0.03	-0.01	0.02	0.01	0.00
ME	-0.02	-0.02	0.00	0.00	0.02	0.00	-0.03	-0.02	0.01	0.01	-0.02	0.02	-0.03	-0.03	-0.02	-0.02
MI	-0.21	-0.28	-0.05	-0.14	-0.15	-0.25	-0.28	-0.26	-0.16	-0.04	-0.17	-0.16	-0.03	-0.27	-0.18	-0.01
MN	-0.09	-0.11	-0.03	-0.04	-0.03	-0.08	-0.12	-0.10	-0.04	-0.01	-0.08	-0.03	-0.05	-0.12	-0.08	-0.03
MO	-0.15	-0.20	-0.06	-0.10	-0.11	-0.17	-0.19	-0.18	-0.10	-0.04	-0.12	-0.11	-0.04	-0.18	-0.13	-0.02
MS	0.08	0.08	0.01	0.04	0.02	0.08	0.09	0.09	0.05	0.04	0.05	0.02	0.05	0.11	0.06	0.04
MT	0.07	0.11	0.03	0.03	0.02	0.07	0.12	0.10	0.03	0.02	0.06	0.02	0.05	0.11	0.06	0.03
NC	0.00	0.03	0.00	0.02	0.04	0.03	0.01	0.02	0.03	0.02	0.00	0.04	-0.01	0.02	0.00	0.00
ND	-0.09	-0.11	-0.03	-0.04	-0.03	-0.08	-0.12	-0.10	-0.04	-0.01	-0.08	-0.03	-0.05	-0.12	-0.08	-0.03
NE	-0.09	-0.11	-0.03	-0.04	-0.03	-0.08	-0.12	-0.10	-0.04	-0.01	-0.08	-0.03	-0.05	-0.12	-0.08	-0.03
NH	-0.02	-0.02	0.00	0.00	0.02	0.00	-0.03	-0.02	0.01	0.01	-0.02	0.02	-0.03	-0.03	-0.02	-0.02
NJ	0.01	0.02	0.00	0.02	0.03	0.04	0.01	0.02	0.04	0.02	0.01	0.03	-0.01	0.02	0.01	0.00
NM	0.03	0.02	0.01	0.02	0.02	0.04	0.02	0.03	0.05	0.02	0.02	0.02	0.02	0.03	0.02	0.02
NV	0.07	0.11	0.03	0.03	0.02	0.07	0.12	0.10	0.03	0.02	0.06	0.02	0.05	0.11	0.06	0.03
NY	0.01	0.04	0.01	0.03	0.05	0.05	0.02	0.03	0.04	0.03	0.01	0.05	-0.02	0.02	0.01	-0.01
OH	0.01	0.01	0.00	0.02	0.03	0.02	0.00	0.01	0.02	0.00	0.01	0.03	-0.03	0.01	0.01	-0.02
OK	-0.01	0.03	-0.01	0.00	0.01	0.01	0.03	0.02	-0.01	-0.01	-0.01	0.01	-0.01	0.02	-0.01	0.00
OR	0.07	0.11	0.03	0.03	0.02	0.07	0.12	0.10	0.03	0.02	0.06	0.02	0.05	0.11	0.06	0.03
PA	0.01	0.02	0.00	0.02	0.03	0.04	0.01	0.02	0.04	0.02	0.01	0.03	-0.01	0.02	0.01	0.00
RI	-0.02	-0.02	0.00	0.00	0.02	0.00	-0.03	-0.02	0.01	0.01	-0.02	0.02	-0.03	-0.03	-0.02	-0.02
SC	0.00	0.03	0.00	0.02	0.04	0.03	0.01	0.02	0.03	0.02	0.00	0.04	-0.01	0.02	0.00	0.00
SD	-0.09	-0.11	-0.03	-0.04	-0.03	-0.08	-0.12	-0.10	-0.04	-0.01	-0.08	-0.03	-0.05	-0.12	-0.08	-0.03
TN	0.03	0.05	0.01	0.03	0.04	0.06	0.04	0.05	0.05	0.02	0.02	0.04	-0.01	0.05	0.02	0.00
TX	0.00	-0.02	0.00	-0.02	-0.04	-0.02	-0.01	-0.01	-0.03	0.01	0.00	-0.04	0.04	0.01	0.00	0.04
UT	0.07	0.11	0.03	0.03	0.02	0.07	0.12	0.10	0.03	0.02	0.06	0.02	0.05	0.11	0.06	0.03
VA	0.00	0.03	0.00	0.02	0.04	0.03	0.01	0.02	0.03	0.02	0.00	0.04	-0.01	0.02	0.00	0.00
VT	-0.02	-0.02	0.00	0.00	0.02	0.00	-0.03	-0.02	0.01	0.01	-0.02	0.02	-0.03	-0.03	-0.02	-0.02
WA	0.07	0.11	0.03	0.03	0.02	0.07	0.12	0.10	0.03	0.02	0.06	0.02	0.05	0.11	0.06	0.03
WI	0.02	-0.02	0.00	-0.01	-0.04	-0.01	0.01	0.00	-0.02	0.00	0.01	-0.04	0.05	0.02	0.01	0.04
WV	0.01	0.01	0.00	0.02	0.03	0.02	0.00	0.01	0.02	0.00	0.01	0.03	-0.03	0.01	0.01	-0.02
WY	0.07	0.11	0.03	0.03	0.02	0.07	0.12	0.10	0.03	0.02	0.06	0.02	0.05	0.11	0.06	0.03