Processing Power: The Effect of Data Centers On Wholesale Electricity Markets

Owen Kay ⓡ Robert Reaser ⓡ Reid Taylor*

3rd May, 2026

Abstract

Artificial-intelligence-driven data centers are reversing two decades of flat U.S. electricity demand and have generated questions about how this growth will impact electricity prices. We quantify this effect using an hourly, unit-level least-cost dispatch model covering wholesale electricity markets in the continental United States. We find that existing data centers have already increased wholesale prices by 2 to 6% on average nationwide, with substantially larger effects in regions hosting major data center corridors. Extending the model through 2028, we show that if proposed construction proceeds under high-utilization scenarios, wholesale prices could rise dramatically (50%), while more moderate build-out yields smaller (20%) but still meaningful effects. Impacts vary due to utilization and build-out assumptions. Finally, we use the model to address several policy discussions including optimal data center siting decisions and renewable build-out uncertainty.

JEL code: L94, P18, Q41, Q42, Q48
Keywords: electricity prices, energy, data centers, artificial intelligence

*Kay: Federal Reserve Bank of Dallas (email: owen.a.kay@gmail.com). Reaser: University of California Davis (email: rreaser@ucdavis.edu). Taylor: Federal Reserve Bank of Dallas (email: reidbtaylor@gmail.com). ⓡ indicates that author order was randomized and all authors contributed equally. We would like to thank Jim Bushnell, Catie Hausman, Lutz Kilian, Erich Muehlegger, and David Rapson for helpful comments as well as seminar participants at CAISO, the Dallas Fed, UC Davis, and UC Santa Barbara. The views expressed here should not be interpreted as reflecting the opinions of the Federal Reserve Bank of Dallas, the Federal Reserve Board or of any other person associated with the Federal Reserve System.

1 Introduction

After nearly two decades of flat electricity consumption in the United States, load growth has re-emerged as a defining feature of power markets. A central driver of this reversal is the rapid expansion of data centers, fueled by recent advances in artificial intelligence (AI). Unlike traditional sources of demand growth, such as households, commercial buildings, or gradual electrification, data centers arrive in large, discrete increments, cluster geographically, and operate at high utilization rates. Their expansion therefore represents a qualitatively different form of load growth: concentrated rather than diffuse, persistent rather than transitory, and fast relative to the pace of generation and transmission investment.

These features matter because electricity markets are characterized by non-linear and steep short-run supply curves (Borenstein 2002). Generator outages, transmission constraints, fuel availability, and limited interconnection capacity imply that incremental demand is often met on sharply rising portions of the merit order. When supply is elastic, new load primarily increases production with modest price effects. When supply is tight, even moderate demand increases can produce large changes in generation costs, prices, and emissions. While these mechanisms are well documented in the context of weather shocks and outages, far less is known about how they operate in response to persistent, geographically concentrated demand growth such as that generated by data centers.

This paper quantifies the incremental effect of data center expansion on wholesale electricity market outcomes in the United States. We focus on three main equilibrium objects: generation costs, competitive wholesale prices, and carbon emissions. Our approach is structural. We construct hourly supply curves for conventional generators using unit-level data and estimates of capacity, outages, fuel costs, operating costs, and emissions. We then solve a least-cost dispatch model that clears regional electricity markets hour by hour. We combine this framework with detailed data on the location, timing, and power capacity of data centers to compare market outcomes to various counterfactual data center utilization scenarios.

The first part of the paper is retrospective and isolates the marginal contribution of data centers to market outcomes over the recent period of rapid expansion from 2021-2025. For each hour and region, we compute equilibrium dispatched quantities and prices under observed conditions and under alternative "no data center" counterfactuals. This comparison yields estimates of the additional generation costs, price impacts, and emissions attributable to data center demand, holding fixed contemporaneous supply conditions and inter-regional power flows. Because data center utilization is not directly observed, we evaluate a range of capacity and utilization scenarios, allowing impacts to vary with both the scale and temporal profile of computing load.

In the second part of the paper we analyze how a number of different scenarios for the data center build-out would impact wholesale electricity markets in the future. Using pre–AI-boom demand projections together with announced generation additions and retirements, we construct hourly demand and supply paths under alternative potential data center growth scenarios. These simulations characterize how future prices and emissions depend on the interaction between data center deployment, utilization patterns, and the pace of supply expansion. Importantly, these forward looking exercises are disciplined by the same dispatch framework used in the retrospective analysis, linking projected impacts directly to short-run market responses. We find that in our high-capacity demand scenario, monthly average wholesale prices increase by \$10–\$40, nationwide. Mid- and low-use demand scenarios still see sizable effects from \$5 to \$20.

Our results highlight three main findings. First, realized data center growth has already increased generation costs and wholesale prices in regions facing tight supply constraints, with effects that are nonlinear in system load. Price impacts are largest when data center demand coincides with periods of scarcity or congestion, reflecting the convexity of short-run supply. Second, these price effects are accompanied by shifts in dispatch toward higher-cost and, in many cases, higher-emissions generation, implying that data center expansion has environmental consequences even absent changes in average emissions rates. Third, counterfactual scenarios indicate that future impacts are highly uncertain and effects could be substantially larger if projected data center capacity comes online faster than new generation or transmission, particularly renewable energy

sources which currently constitute the majority of proposed capacity additions.

We use our results to discuss a number of contemporary policy issues. First, we discuss how the wholesale price effects can pass through to end-use retail customers. Our empirical approach simulates the effects of variable costs due to data centers, but residential customer bills are used to recover both variable and fixed costs. The effect of data centers on customer bills is theoretically ambiguous as large increases in demand could decrease average fixed costs (Borenstein 2025). However, the data center expansion likely will require large investments in transmission and distribution upgrades to the grid and even if average fixed costs decrease, the variable cost component of electricity bills is unambiguously increasing. We calculate the conditions under which new load can decease customer bills, showing that it requires the share of customer bills coming from fixed costs to be very high. We next explore how market integration through increased transmission is able to lower the effect of data centers by allowing cheaper, idle generation to serve demand that was previously in a separate market. Lastly, as an alternative, we explore how the spatial reallocation of data center compute can act as virtual transmission by comparing simulated outcomes to an optimal spatial allocation of compute that minimizes hourly costs of generation.

We contribute to a growing literature on the economic and environmental consequences of electricity-intensive computing infrastructure. As projections for data centers have ballooned, a number of technical reports have estimated the energy demands associated with this construction boom (Shehabi et al. 2024; Green et al. 2024). However, there have been relatively few academic studies on the economic impacts of these changes. Bogmans et al. (2025) provide quantitative projections of the effect of AI-driven data center growth on electricity prices and carbon emissions at economy-wide scale using a multi-country computational general equilibrium model. Similarly, recent work has begun to document localized wholesale price effects of data center expansion in major corridors (Mamkhezri et al. 2025; Feher et al. 2025), while related research on cryptocurrency mining, an extreme form of high-utilization computing load, finds increases in retail electricity prices with only partial fiscal offsets (Benetton et al. 2023). Muller (2026) measures impact of data centers on local pollution and greenhouse gas emissions using regional average

emissions rates of power generation. Other work by Gargano and Giacoletti (2025) studies the political economy of attracting data centers, showing that state incentives meaningfully shift data center siting and investment, without commensurate increases in local tech employment.

In contrast to much of this earlier work, our approach constructs a detailed model of electricity supply at a precise geographic and temporal scale. This approach allows us to appropriately capture the non-linear price response to changes in demand that occurs when particular markets or times are constrained. In the most closely related work to our own, Wade et al. (2025) use a capacity expansions model to forecast long run effect of rising demand from data centers and crypto mining in U.S. power markets. However, they lack detailed geographic and temporal data on data centers and instead assume fixed growth rates in compute from demand. In contrast, we use detailed site-specific data on existing and proposed data centers to model the wholesale electricity market impacts at a detailed geographic and temporal level through 2028.

More broadly, our analysis underscores that the consequences of demand growth depend not only on how much electricity is consumed, but on where and when that consumption occurs and on the availability of marginal supply. Recognizing this point, a growing literature has emphasized that if demand is able to flexibly avoid congested hours and locations, impacts on wholesale markets can be mitigated (Norris et al. 2025). This literature has mostly focused on temporal flexibility and how this flexibility maybe impact which generation comes online and emissions profiles (Knittel et al. 2025; Ross and Ewing 2026). In contrast, our analysis of siting decision documents how geographic flexibility in where compute occurs can provide system benefits. As data centers continue to expand, understanding these interactions will be essential for infrastructure planning, market design, and climate policy. By quantifying both realized impacts and plausible future trajectories, this paper provides an empirical foundation for those discussions.

2 Data

2.1 Electricity Supply Data

Our primary data source for electricity supply is the Environmental Protection Agency's (EPA) Continuous Emission Monitoring System (CEMS), which reports hourly electricity production and fuel use at nearly every conventional generator in the U.S.[1] CEMS reports generator *gross* electricity production, which we convert into net production using EIA-923 data to calculate observed plant-level ratios. For each generation unit, we observe characteristics including the technology and fuel type, precise geographic location, emissions of regulated air pollutants, and coverage by environmental regulations. We calculate maximum generator capacity as the 99th percentile of observed generation. Generators are grouped into regions based on the EPA's eGRID subregions.

We closely follow Hausman (2025) and Ham et al. (2025) to construct hourly market supply curves. Generator-specific marginal costs are calculated as the sum of fuel costs, variable operations and maintenance costs, and pollution permit prices. Specifically, for generator i in hour t, the marginal cost is given by

$$MC_{it} = HeatRate_i * FuelCost_{it} + VOM_i + Permit_{it}. \quad (1)$$

The heat rate for generator i, $HeatRate_i$, is calculated from the observed data as the ratio of total fuel consumption to annual net generation.[2] For coal- and coke-fired generators, we use data from the EIA on state specific monthly fuel costs paid by power plants. For natural gas and oil-fired generators, we use daily fuel spot prices (Henry Hub and West Texas Intermediate) and apply a state-by-month specific markup, computed from EIA data on average fuel prices paid by

[1] CEMS comprises the universe of coal, natural gas, oil, and other combustible fuel generators with capacity over 25 MW, but not nuclear or renewable generators. Following Ham et al. (2025), we drop commercial, industrial, and cogeneration units whose main function is not selling into electricity markets.

[2] We censor anomalous heat rates below 4 mmBtu per MWh and above 50 mmBtu per MWh.

power plants. We observe environmental trading program permit prices from the EPA and the Regional Greenhouse Gas Initiative (RGGI). Technology-specific O&M costs are calibrated from the EIA based on Hausman (2025). Furthermore, power plants are often offline for either routine maintenance or unexpected outages. We follow Hausman (2025); Ham et al. (2025) and uniformly derate generator capacity by the planned outage rate and stochastically apply unplanned outages to individual hours. All dollar values are inflation adjusted using the total CPI index to 2024 values.

Table 1 provides summary statistics for the thermal generator fleet in 2025. Coal- and natural gas-fired power plants make up the vast majority of the thermal generation fleet. The range of marginal costs for natural gas units is much larger than for coal-fired units, reflecting both the volatility in gas prices and the large variance in heat rates (efficiency) coming from different technology types (combined cycle vs simple cycle). There is also a significant amount of oil-fired generation capacity but these units tend to be much higher marginal cost and are often among the last generators to be dispatched.

Table 1: Thermal Generator Characteristics Statistics

Fuel	N Units	Avg Capacity	Min Capacity	Max Capacity	Mean MC	Min MC	Max MC
Coal	368	456.6	24.6	1338.6	31.3	13.4	82.3
Coke	5	156.5	19.4	288.9	38.2	35.6	42.5
NG	2761	159.0	2.6	955.2	47.2	14.3	557.8
Oil	234	53.2	1.9	551.9	356.4	115.0	856.5

Notes: This table summarizes the capacity and marginal costs of the 3368 thermal generation units active in the lower 48 states in 2025. The unit of observation is the generator unit and the columns report the average, minimum and maximum capacity and marginal cost for unit of each primary fuel type in 2025.

2.2 Data Centers

We obtain data on the characteristics of existing and planned data centers including the precise geographic location, peak power demand (MW), operational dates, capital cost, square footage, and site acreage from the Cleanview data center tracker. The data set covers all categories of data centers.

For our model of electricity markets, we require a measure of hourly data center power de-

mand. There are, however, three sources of uncertainty in the power demand for data centers. First, not all data centers release public information on the size of their infrastructure. Approximately 67% of data centers in our data have a reported power capacity. For those with missing power data, we impute capacity using the year of first operation and site characteristics. Second, while the maximum power demand may be known, actual utilization is unknown for all data centers. Although many data centers publicly state their intent for 100% utilization, industry reports show real-world utilization is much lower and varies by data center type. This motivates our analysis to test the sensitivity of our results to various utilization assumptions. Lastly, some data centers have installed on-site "behind-the-meter" backup power generation and can produce their own electricity. Information on the size and utilization of "behind-the-meter" sources is not publicly reported. In our various utilization scenarios, we are agnostic as to whether the reduced utilization is due to lack of computing needs or a switch to on-site generation as they have the same impact on the electricity grid.

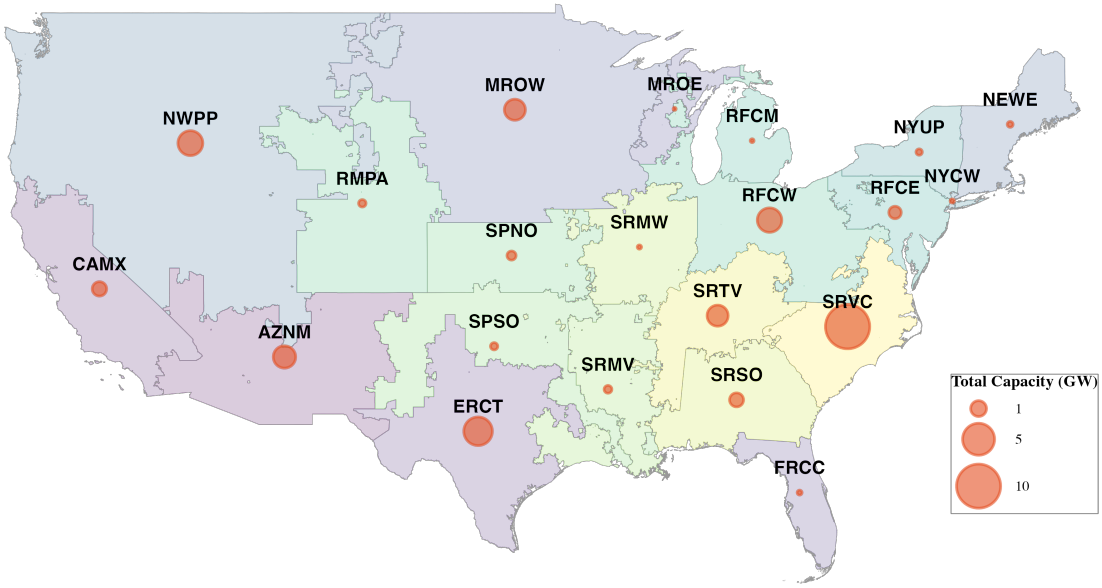
We capture this uncertainty by constructing four data center demand profiles that allow us to analyze the effect on prices across a variety of potential scenarios. We vary both the power capacity (the maximum power draw at a point in time) and the utilization rate (the percentage of maximum used in a given hour). In the Full-Capacity scenario, we assume 100% of installed power capacity is used 100% of the time. For the reasons noted above, this scenario is highly unlikely, but provides a useful upper bound. We leverage data from Bloomberg/DC Byte, which reports average annual utilization rates and market share by data center type as reported by industry participants. We scale data center capacity and utilization by these rates to construct two plausible scenarios. First, we allocate a percentage of power equal to the market share for training data centers to every hour as these have a 100% utilization rate. Next, we allocate the remaining portion according to the weighted average utilization rate across all other data center types. In our Mid-Range scenario, we assign this power equally across all hours of the day. In our Mid-Range-Peak scenario, data centers only operate during business hours, operating at full capacity the required number of hours to meet the observed utilization rate every day, centered around 1PM. Lastly, our Low-Capacity

scenario reduces the capacity of data centers, assuming 50% less power in each hour than in the Mid-Range scenario.

While data centers are siting across the country, Figure 1 shows the geographic distribution and concentration of operational data center capacity in 2025, and Figure 2 shows operational and proposed data center capacity as of 2030. To date, Virginia has led the way with the most operational data center capacity in the country. However, by 2030, Texas (and thus the ERCOT grid) is slated to be the leader in data center power demand.

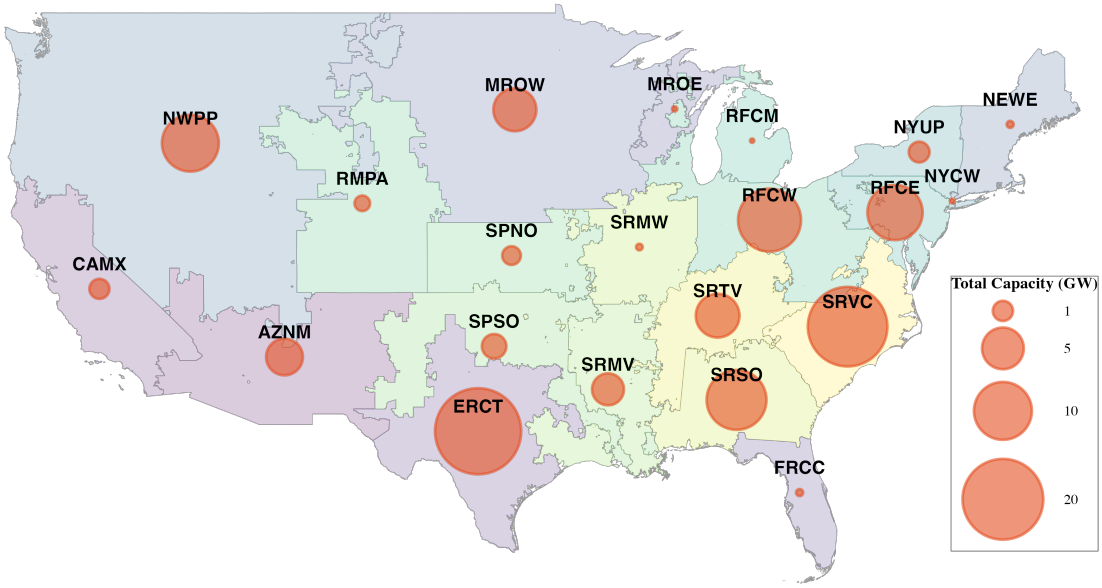
Data centers consider a combination of factors when deciding where to locate. Electricity access is a primary consideration. Grid interconnection queues can often require projects to wait years for access to connect, driving data centers to areas that will allow them to connect quickly. While cheaper electricity rates can lower costs for data centers, they do not exclusively locate in grids with the lowest prices. Using data from the Census Bureau American Community Survey and the EIA, we report demographic variables for the counties where data center siting occurs in Appendix Table A1. Data centers locate along major fiber optic cable corridors and often near major urban areas to reduce latency and benefit from skilled labor markets. This pattern is evident when comparing differences between counties with and without data centers. As noted before, counties with data centers have marginally higher electricity rates than areas without.

Figure 1: Installed Data Center Capacity: 2025



Notes: Map plots the operational data center power capacity as of the end of 2025 by eGRID subregion. Data are sourced from Cleanview. Larger circles represent more cumulative capacity in the subregion.

Figure 2: Installed and Planned Data Center Capacity: 2030



Notes: Map plots the operational data center power capacity as of the end of 2030 by eGRID subregion. Data are sourced from Cleanview. Larger circles represent more cumulative capacity in the subregion.

3 Structural Dispatch Model

To analyze both the ex post period and the forward looking scenarios, we construct a short-run competitive dispatch benchmark for each subregion. In each hour, conditional on residual demand to be served by inframarginal generators, available unit-level capacity, and unit-level marginal costs, we solve a linear program that minimizes the total cost of meeting load within the subregion. Formally, for each region r and hour t, we solve

$$\min_{q_{it}} \quad \sum_{i \in I_r} mc_{it} q_{it} \tag{2}$$

$$\text{s.t.} \quad \sum_{i \in I_r} q_{it} = d_{rt} \quad \forall r,t \tag{3}$$

$$0 \leq q_{it} \leq K_{it} \quad \forall i \in I_r, t. \tag{4}$$

Here, I_r denotes the set of conventional generating units in region r, q_{it} is output from unit i in hour t, mc_{it} is the marginal cost of unit i in hour t, and d_{rt} is residual demand in region r and hour t after accounting for non-thermal generation. Generation from each unit is bounded above by available derated capacity K_{it}. Available capacity varies at the hourly level in the model. We first apply monthly derating factors to account for expected outages and then incorporate a stochastic hourly outage component to capture unexpected outages, so that in some hours a unit may be unavailable and $K_{it} = 0$.

Solving this problem yields two objects of interest: optimal dispatch quantities, q^*_{it}, and the shadow value on the power-balance constraint, λ^*_{rt}. We interpret λ^*_{rt} as the marginal cost of serving an additional MWh of residual demand in region r at hour t. Because the model is based on unit-level marginal costs rather than strategic offers and abstracts from non-convex operating constraints such as start-up costs and ramping constraints, λ^*_{rt} should be interpreted as a competitive marginal-cost benchmark.

This interpretation is also relevant outside organized wholesale markets. In regions served primarily by vertically integrated utilities, there is often no directly observed wholesale market price analogous to that in RTOs and ISOs. In those regions, λ^*_{rt} remains economically meaningful as the model-implied marginal cost of serving an additional MWh of load. For expositional simplicity, we refer to λ^*_{rt} throughout as the regional price.

Our counterfactual is best viewed as a short-run partial-equilibrium exercise that isolates how the marginal cost of serving load changes when data center demand is added or removed, condi-

tional on realized non-thermal generation, generator availability, and estimated marginal costs. The framework does not recover a full market equilibrium for the interconnected U.S. power system, as it abstracts from endogenous interregional trade, strategic bidding, and non-convex operating constraints. Accordingly, the resulting regional price should be interpreted as a model-based benchmark under observed system conditions rather than as a literal prediction of the realized hourly market-clearing price.

4 Ex Post Analysis

In this section, we use the dispatch framework from Section 3 to quantify the contribution of data center load to generation costs, model-implied regional prices, and emissions for 2021-2025. Throughout, these estimates should be interpreted as short-run partial-equilibrium counterfactuals: they measure how market outcomes change when data center demand is added or removed, holding fixed contemporaneous supply conditions, non-thermal generation, and inter-regional power flows.

Each of our alternatives is compared to the reference case where we set net load for conventional generation equal to observed thermal generation in each region. For the data center utilization scenarios, we subtract implied data center demand from observed net generation in each hour, to produce a no-data-center residual demand scenario for each counterfactual. We then re-solve the dispatch model for q_{it} and λ_{rt} without data center demand.

By setting our baseline residual demand for thermal generation equal to observed thermal generation in each region, we allow thermal generation to be re-optimized but hold other source of supply and demand fixed. Embedded in this modeling decision is an assumption that other sources of supply such as nuclear, wind, solar, and hydro are always inframarginal. This is consistent with very low marginal costs for these sources of supply.

In addition, we are implicitly fixing the net interchange between regions to their historic levels. Observed generation in each region is equal to regional electricity consumption plus net exports. For the counterfactual scenario that includes data center demand, this simply ensures

that net exports between the various regions match observed net exports. In our no-data-center counterfactuals, we subtract the hourly assumed data center load from observed generation. Thus, we also fix transmission flows between regions to their observed levels, even in the counterfactual without data center demand. This implies that the marginal data center electron is always generated within the same region. In reality, if a neighboring region had cheaper generation available and transmission lines between regions were not at capacity, we would expect incremental data center load to be partially met by the neighboring region.

These considerations motivate our use of eGRID subregions to define our geographic markets. EPA's methodology for assigning plants to eGRID regions is intended to limit imports and exports within an aggregated area in order to produce regional emissions rates that reflect the electricity consumption within that geographic area. To the extent that the eGRID subregion boundaries successfully limit bias introduced by imports and exports, our main results can be interpreted as the additional costs, emissions, and price impacts of data center load. Because we fix historic trade flows, our model does not capture inter-regional re-dispatch that would occur in a no-data-center counterfactual. In that respect, our cost estimates will understate the true impacts of data center load. The direction of the bias on emissions and local price impacts is ambiguous. Changes to emissions depend on the regional generation mix. Local price impacts would be biased up in cases where a region is a net-importer and biased down for net-exporters.

We then compute counterfactual dispatch under "no-data-center" counterfactual demand scenarios. To do so, we subtract data center electricity demand from the measure of total conventional generator electricity demand used in the base case. We only observe capacities for existing data centers, not utilization rates, so we calculate total data center electricity demand under a number of alternative assumptions of utilization rates described in Section 2. For each "no-data-center" demand scenario, we solve the dispatch model for generator production levels, q_{it}^{-DC}, and wholesale prices, λ_{rt}^{-DC}, at each hour had there been no data center demand.

The model is able to recover market outcomes accurately. In Appendix Figure A9, we show a plot of observed vs. simulated dispatch at the plant-hour level for a random subset of the sample.

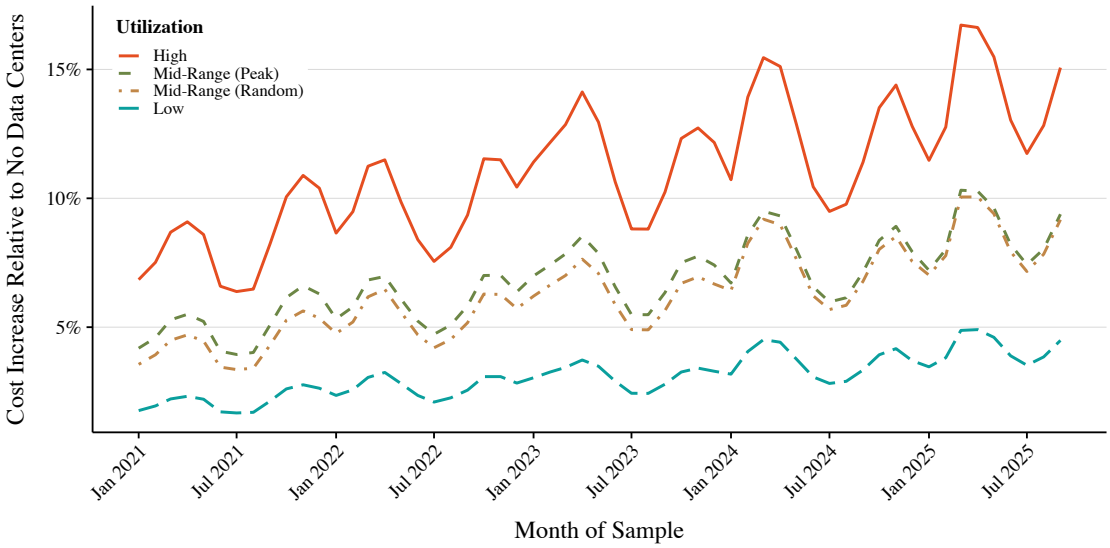
There is a mass of observations along the 45-degree line, indicating that our simulated hourly dispatch instructions closely match actual market outcomes. In Appendix Table A3, we calculate correlations between the simulated and observed generation. For hourly generation, there is a correlation coefficient of .631, suggesting a strong positive relationship. Due to our application of stochastic outages and abstraction from transmission congestion, this number is expected to be less than 1, and is in line with similar applications in the literature (Hausman 2025; Ham et al. 2025). When we aggregate to month and yearly levels of generation, we see even stronger correlations between observed and simulated, reaching .863 for the unit-by-year level. This is further evidence that stochastic outages play a role in the hourly-level correlations.

4.1 Additional Generation Costs

We begin by calculating the difference in total generation costs between the simulated baseline and each no-data-center counterfactual, $\sum_i mc_{it} q_{it} - \sum_i mc_{it} q_{it}^{-DC}$. We compare the simulated no-data-center counterfactuals to the simulated baseline, rather than to observed generation, so that the resulting differences reflect only changes in dispatch induced by removing data center load.[3]

[3]Generator-specific production in the simulated baseline may differ from observed production for several reasons, including mismatch between simulated and realized outages, the omission of ramping constraints, and imperfect representation of the underlying transmission network.

Figure 3: Percentage Increase in Monthly Generation Costs
Attributed to Data Center Load, 2021–2025Q3



Notes: Graph shows the load-weighted average monthly incremental thermal generation costs as a percentage of no-data-center generation costs by various data center utilization scenario.

Figure 3 reports the percentage increase in simulated monthly generation costs in the baseline relative to each no-data-center counterfactual. In the high-utilization scenario, in which all operational data centers are assumed to run at full power in every hour, we estimate that by 2025 data center load increased generation costs by roughly 10 to 15% nationwide, depending on the season. At the other end of the scenario range, the low-utilization case implies cost increases of at most about 5%. These estimates should be interpreted as a range across utilization assumptions rather than as formal bounds. More generally, except in hours with negative marginal prices, additional load mechanically increases variable generation costs even if its effect on the model-implied price is smaller or more heterogeneous.

Focusing on the mid-range scenarios, which are calibrated to observed utilization rates, we estimate additional costs of 5% at the beginning of our sample, increasing to a peak of 10% in spring of 2025. The scenario which concentrates data center demand in peak business hours yields generation costs that are, at times, an additional 1% higher than the scenario in which usage is spread evenly throughout the day.

The estimates also reveal a notable seasonal pattern. Cost increases in the summer and winter

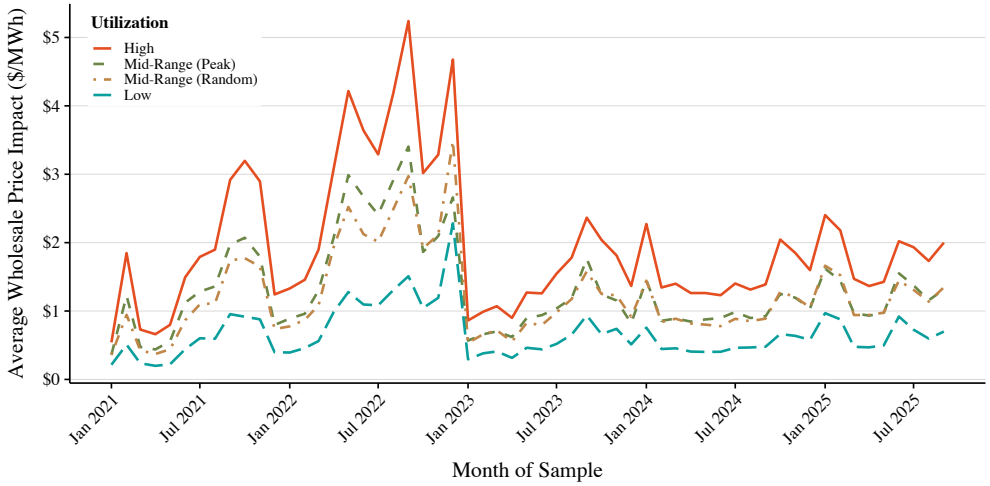
tend to be several percentage points lower than in the shoulder seasons. A likely explanation is that generators commonly schedule maintenance outages in the spring and fall, when demand is typically lower because of milder weather. As a result, available thermal capacity is reduced in those months, and the relevant portion of the thermal supply curve becomes steeper. As data center load expands over the sample, the market therefore clears on a steeper portion of the supply curve in the shoulder seasons than it would in the absence of that load raising costs by more for a similar shift in demand.

4.2 Competitive Wholesale Price Effects

We next turn to the competitive benchmark price in the wholesale market. While cost of generation is a relevant measure for the areas of the country covered by vertically integrated utilities, for the majority of the country covered by competitive wholesale electricity markets, the market price is relevant both for generator profits and the price paid by utilities to procure electricity for retail customers.

Figure 4 contains our estimates of the nationwide load-weighted average monthly price effects of data centers under the same four data center demand scenarios. Similar to our characterization of additional costs, we use the difference between the baseline competitive wholesale price using observed net load and our counterfactual scenarios. By late 2025Q3, the marginal generator dispatched to meet electricity demand was approximately $0.50 to $2.00 more expensive, on average, than what would have been dispatched in a the counterfactual world without data center load.

Figure 4: Average Monthly Competitive Wholesale Price Impact, 2021–2025Q3



Notes: Graph shows the load-weighted average monthly simulated market price impact of data centers by data center utilization scenario.

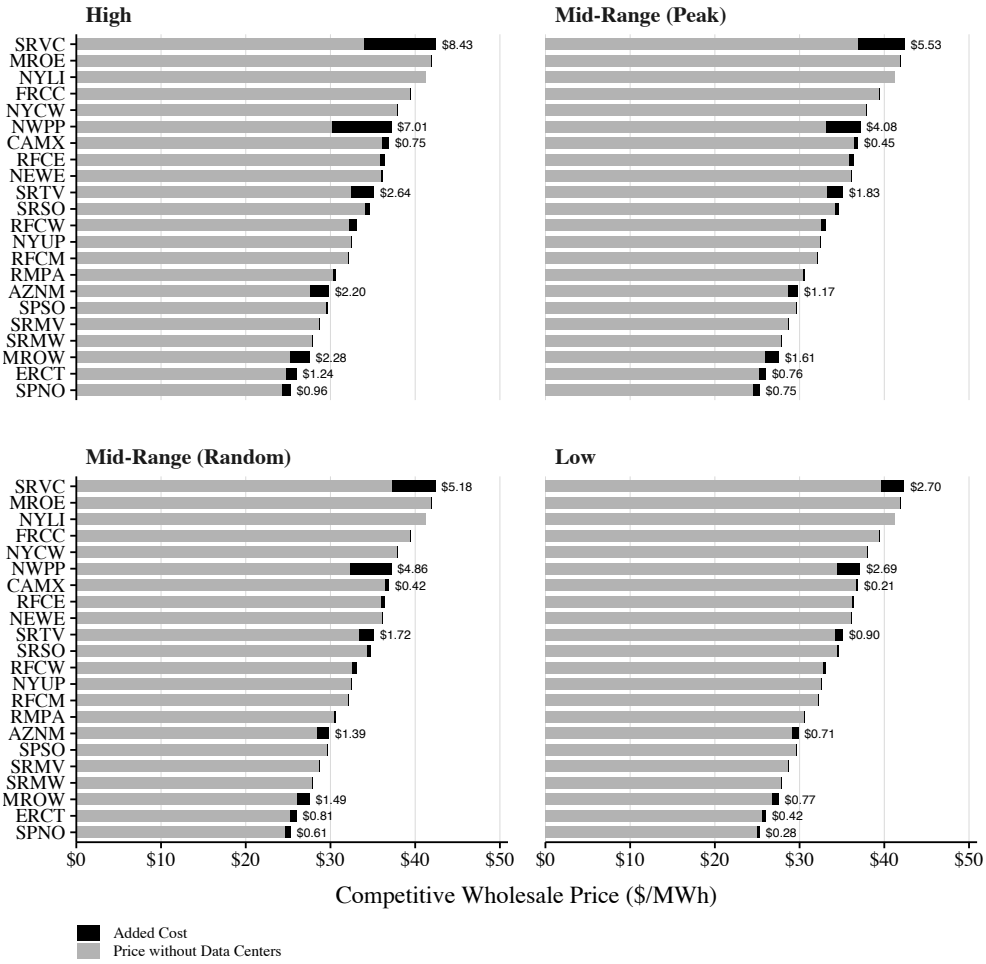
The highest estimated price effects occurred during the summer of 2022. Large spikes to natural gas prices caused by the Russian invasion of Ukraine increased wholesale electricity prices as natural gas generators are often the marginal generator, and set the market price. Our estimates show that during this period, data center demand added an additional $1 to $5 to the nationwide wholesale price, depending on the utilization assumption of data centers.

The distribution of data center load across the U.S. and the heterogeneity in regional power plant fuel composition means that there is substantial variation in results by region. In Figure 5 we show the average price impacts by eGRID subregion from 2023 to 2025Q3. The largest effects are seen in SRVC which is home to "Data Center Alley" in northern Virginia, the densest collection of data centers in the country. Northwest Power Pool (NWPP), which cover the majority of the pacific northwest also has significant effects from data centers. Our mid-range scenarios indicate data centers were responsible for a $5 increase in competitive wholesale prices in SRVC, a greater than 10% increase, and a $4 increase in NWPP.

ERCOT (ERCT), which is the grid covering the vast majority of Texas, has very little price effect despite containing the second highest amount of data centers behind SRVC. Data centers have added less than $1 to wholesale prices, likely driven by nation-leading supply build out. We

also find that data centers had little impact in the subregion covering most of California, CAMX. While Californians face the highest average retail electricity prices in the nation during this period, even our upper-bound estimates yield an impact of less than $0.75.

Figure 5: Average Regional Competitive Wholesale Price Impact, 2024-2025Q3



Notes: Graph shows the load-weighted average simulated market price impact of data centers by data center utilization scenario and eGRID subregion.

Our model allows us to decompose the welfare impacts of data centers associated with our modeled generation. We calculate the total competitive wholesale payments made to dispatchable generators, their total production costs, and the competitive, inframarginal rents earned. Total payments are $\sum_{rt} \lambda_{rt} Q_{rt}$, where λ_{rt} is the competitive wholesale price of subregion r in hour of sample t, and $Q_{rt} = \sum_i q_{it}$ is total thermal net generation in hour t. Production costs are total variable costs (fuel, variable O&M, emissions permits) summed across all dispatched generators and

hours. Competitive rents equal the difference between wholesale payments and production costs. Computing these three objects in our baseline simulations and our no-data-center counterfactuals allows us to identify the changes driven by data center demand.

Table 2: Data Center Impact on Payments, Production Costs, and Competitive Rents, Mid-Range (Peak) Utilization Scenario (Millions of 2024 Dollars), 2021-2025Q3

	Year					
	2021	2022	2023	2024	2025	Cumulative
United States						
Δ Payments	5,836	10,048	6,241	6,096	6,182	34,404
Δ Production Costs	3,723	5,737	4,242	4,057	4,037	21,796
Δ Rents	2,113	4,311	1,999	2,039	2,145	12,607
Restructured Markets						
Δ Payments	2,718	4,159	2,745	2,571	2,817	15,010
Δ Production Costs	1,775	2,461	1,866	1,635	1,701	9,437
Δ Rents	943	1,698	879	936	1,116	5,573
Non-Restructured Markets						
Δ Payments	3,118	5,889	3,495	3,523	3,365	19,390
Δ Production Costs	1,948	3,276	2,376	2,420	2,336	12,356
Δ Rents	1,170	2,613	1,120	1,102	1,029	7,034

Notes: All values in millions of 2024 dollars. Δ rows report the difference between observed (with data centers) and the no-data-centers counterfactual under the Mid-Range (Peak) utilization scenario. Restructured Markets aggregates all ISO/RTO-affiliated facilities; Non-Restructured Markets covers facilities outside organized wholesale markets.

Table 2 presents these three objects computed from our observed dispatch and our no-data-center scenarios.[4] Notably, we find a cumulative welfare transfer from wholesale consumers to suppliers in our mid-peak scenario of $12.6 billion from 2021 to 2025 that was driven by data center demand. The cumulative impact of data centers on production costs totaled over $21 billion. We decompose these results by "Wholesale Markets," regions in the U.S. that have restructured and have a competitive wholesale electricity market, and those that are areas that are still under the traditional, "Vertically Integrated" format, since the interpretations differ with respect to the "Payments" row. In restructured markets, our system price represents the competitive wholesale benchmark. Thus, the values reflect the interpretation above. We find total transfers from wholesale

[4]Past work, such as Borenstein et al. (2002), similarly decompose these welfare objects. Because we only compute a competitive equilibrium in our setting, we make no claims about the market power effects of data centers.

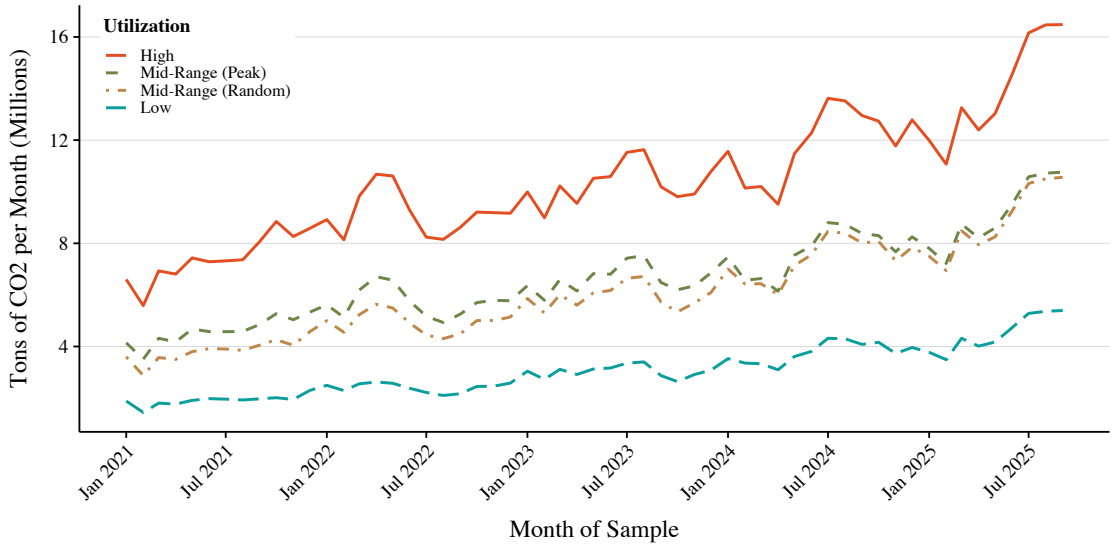
buyers to suppliers of approximately $5.5 billion during our sample period.

In non-restructured, vertically integrated markets, the system price in our model still represents the shadow price of supplying an additional MWh, but this object does not have an explicit mapping to a system price. Thus, we should think of the change in rents in these markets as the gap between the marginal cost of generation and the average cost of all dispatched generation. Under cost-of-service regulation, we would expect these savings to eventually flow back to ratepayers and would not be the persistent windfall that they are in their wholesale counterpart.

4.3 Emissions

From our model results, we observe the additional output required to meet data center load. Using plant specific emissions rates, we can estimate the incremental emissions, measured as tons of CO_2, directly related to data center load growth. Our results, shown in Figure 6, suggest that incremental emissions have been steadily increasing over our sample period as data center load continues to increase, suggesting that data center demand growth has generally outpaced the rate at which new renewable generating capacity has connected to the grid.

Figure 6: Incremental Monthly CO_2 Emissions, 2021–2025Q3

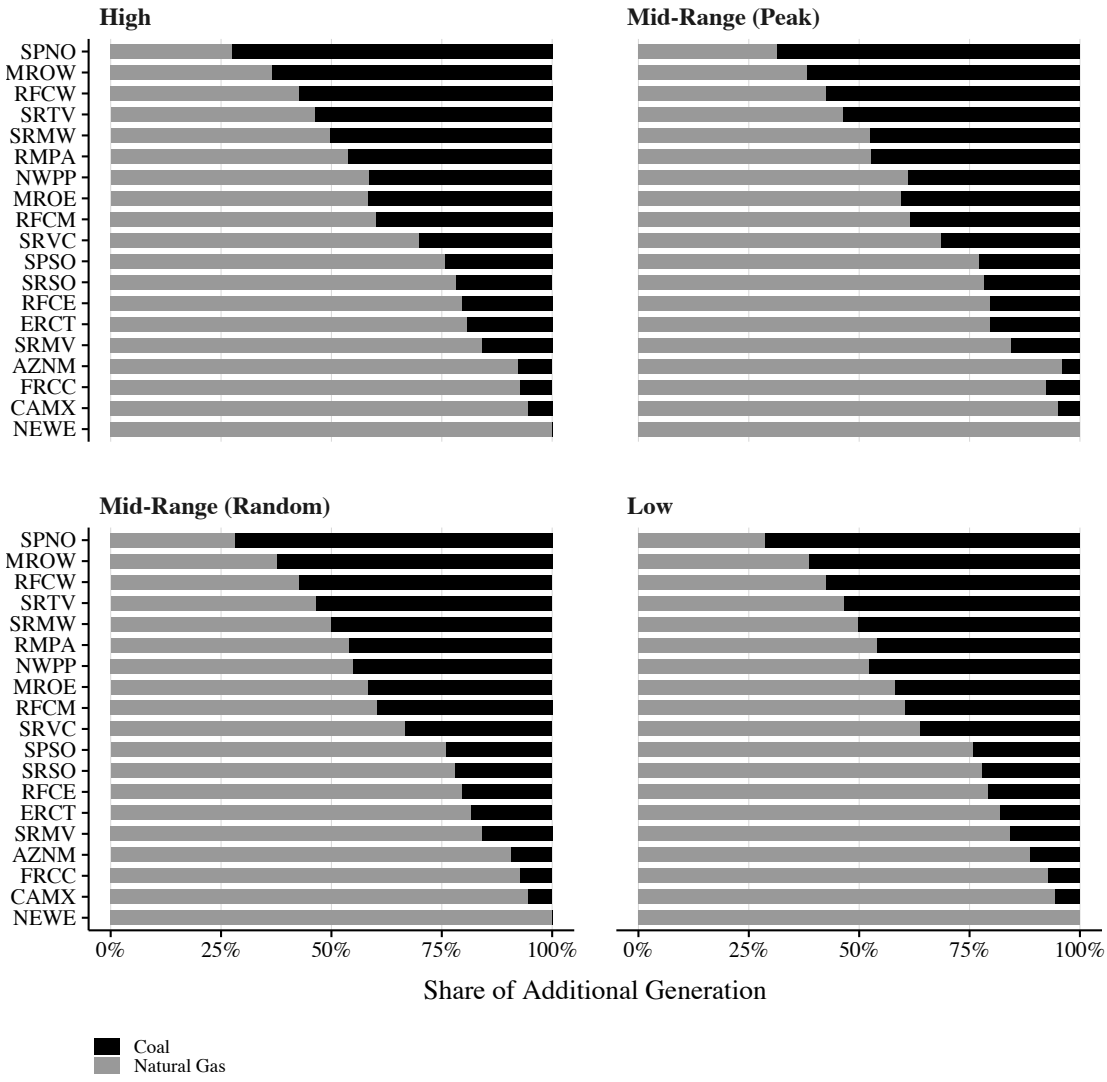


Notes: Graph shows the monthly simulated emissions impact of data centers by data center utilization scenario.

By 2025, data centers were generating between 5 and 16 million additional tons of CO_2 per month. For reference, a moderate social cost of carbon (SCC) estimate of $51, results in external damages ranging from $3 billion to $9.7 billion annually depending on the data center demand assumption. As with the welfare transfer implied by the increase in wholesale prices, this increase in damages must necessarily be weighed against the positive externalities and productivity gains created by data centers to calculate a total effect of data centers, an exercise which is a promising avenue for future research.

It is important to note that not all data centers will have the same emissions impacts. Importantly, the different generation mix across regions and time means that incremental demand is met by generation units of varying efficiencies and technologies. Figure 7 shows the fuels used by the marginal thermal generators to meet data center demand across regions. In New England and California, this is almost entirely natural-gas-fired generation, while in much of the upper Midwest, coal plays an important role.

Figure 7: Share of Incremental Load Met by Coal vs. Natural Gas
By Region, 2024–2025Q3



Notes: Graph shows the generation type breakdown for the incremental generators by data center utilization scenario and eGRID subregion.

5 Forward Looking Analysis

Next we present a forward-looking analysis that quantifies how projected data center expansion changes wholesale market outcomes through 2028. This exercise is complicated by various dimensions of uncertainty. Where and when data centers will be constructed is constrained by regulatory processes and slow-moving electricity interconnections. On the supply side, the timing and

size of power plant additions and retirements are uncertain. Additionally, we must create a credible no-data-center electricity consumption counterfactual to quantify the incremental effect of data centers. Our model is able to handle the wide range of uncertainty and we address each of these in turn by constructing various counterfactual scenarios to test. As such, the exercise should be interpreted as simulating the effects *if* a particular scenario arises, and not assigning a probability to any such outcome.

In general, two results emerge. First, effects are highly state-dependent: under the high-end power usage assumption, data center load drastically raises generation costs and wholesale prices relative to the no-new-data-center benchmark. In fact, demand is so large that in a significant number of hours, there is insufficient available generation to meet the need. The middle and low scenarios result in more moderate, but still economically significant wholesale market effects. Second, the distributional consequences of the build-out are geographically uneven, with larger price effects in regions that are both most exposed to new data center construction and that face supply constraints.

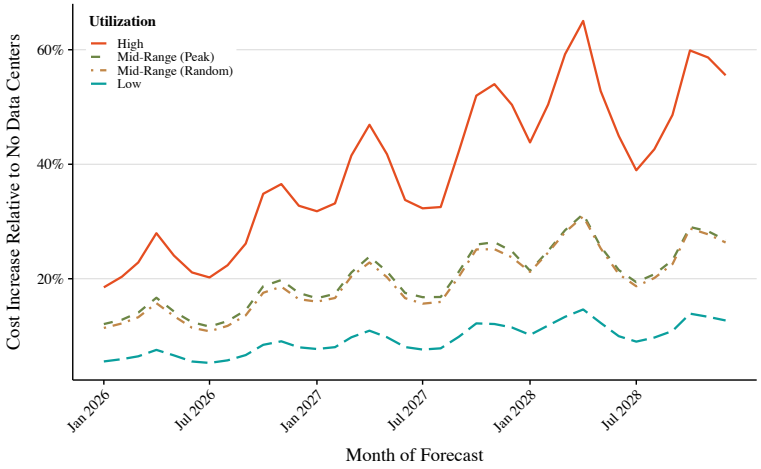
Unlike the ex-post exercise, the forecast requires constructing hourly net load from total forecast demand and non-thermal supply. For each subregion-hour, forecast net load equals projected total demand minus projected output from wind, solar, hydro, storage, and other non-fossil baseload resources as these resources are inframarginal. We then solve the same hourly least-cost thermal dispatch problem outlined in Section 3 to determine the dispatch instructions for thermal generators and market price. Our supply forecast is exogenous and incorporates announced and expected additions, retirements, derates/uprates, outages, and fuel-cost paths from EIA-based inputs. Given the relatively short time frame of our forecast window and interconnection queues for new supply that are greater than 5 years, we believe this assumption is justified.[5] The demand forecast for non-data-center load is derived from the EIA Annual Energy Outlook and data center load for planned locations is reported by Cleanview.[6]

[5] In later robustness checks, we relax this assumption and show the sensitivity of our results to endogenous supply growth and power plant retirements.

[6] Full details on demand and supply forecasts are provided in Appendix D

We evaluate the same four demand scenarios (`high`, `mid_peak`, `mid_rand`, `low`) that differ in effective utilization and intraday load shape. However, to reflect the additional uncertainty from data centers not completing construction or being 'behind the meter', we scale down new data center capacity by 40% of announced for the mid-scenarios. This in turn lowers the low-end case, which is a further 50% reduction. The mid-peak scenario is our preferred central case with high and low scenarios bounding the effects. All effects are reported relative to a no-data-center counterfactual in the same time period.

Figure 8: Projected Percentage Increase in Generation Costs, 2026–2028

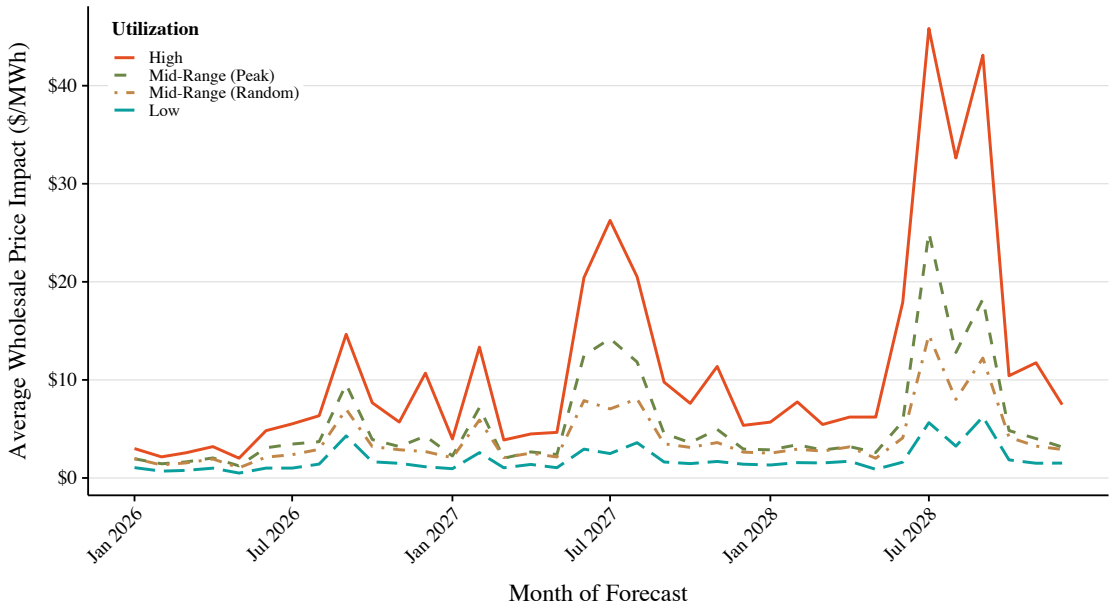


Notes: Graph shows the simulated load-weighted average monthly percentage increase in thermal generation costs by data center utilization scenario.

Figure 8 reports monthly percentage changes in generation cost, and Figure 9 reports monthly wholesale price effects in levels. The very large estimated effects in the high-end case reflect the dramatic announced growth in data centers. If all data centers arrive as planned, and run continuously, generation costs are set to increase by 60% and prices more than $40 higher than a no-data-center counterfactual. While this scenario is highly improbable, it reflects the tail upside risk data centers pose to electricity prices. The mid-scenarios imply more modest, yet still significant, 20–30% increases in generation costs by 2028 with substantial month-to-month variation. Interestingly, the scenario where utilization is concentrated in the peak hours leads to significantly higher market prices when compared to the smoother mid-scenario. Even in the low-end case, generation costs are estimated to grow by 10–15%, with prices rising by close to $10 on average

in some months.

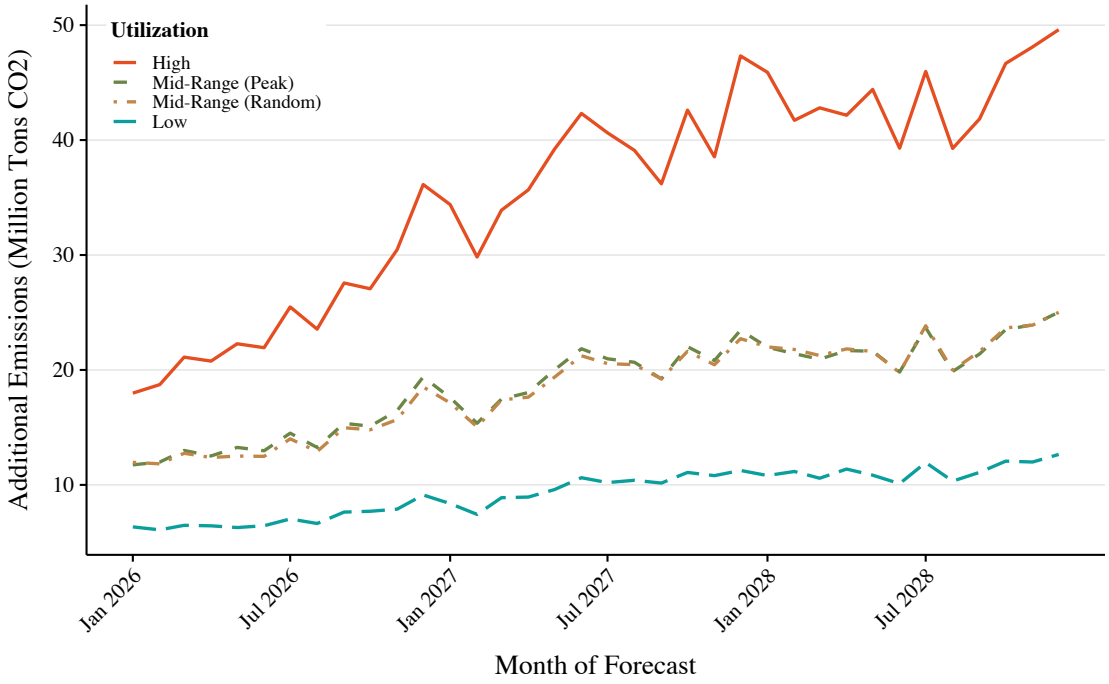
Figure 9: Projected Average Monthly Competitive Wholesale Price Impacts, 2026–2028



Notes: Graph shows the simulated load-weighted average monthly wholesale market price increase by data center utilization scenario.

Figure 10 reports incremental monthly CO_2 emissions from the required additional generation. In the high-demand scenario, data center growth greatly increases emissions, resulting in 45 to 50 million tons of CO_2 per month. In the mid-scenario, emissions increases are still high, reaching over 20 million tons of CO_2 per month by 2028. It is important to note we are measuring grid-generated emissions. Part of the reason we lower proposed capacity is that some data centers may create their own on-site generation 'behind-the-meter'. These sources of energy are often more heavily polluting than on-grid assets, and would be additional to what is shown in Figure 10. The mid-scenario results in around $12 billion annually in additional damages using a social cost of carbon of $51 (or $43 billion annually using a social cost of carbon of $185).

Figure 10: Projected Incremental Monthly CO_2 Emissions, 2026–2028

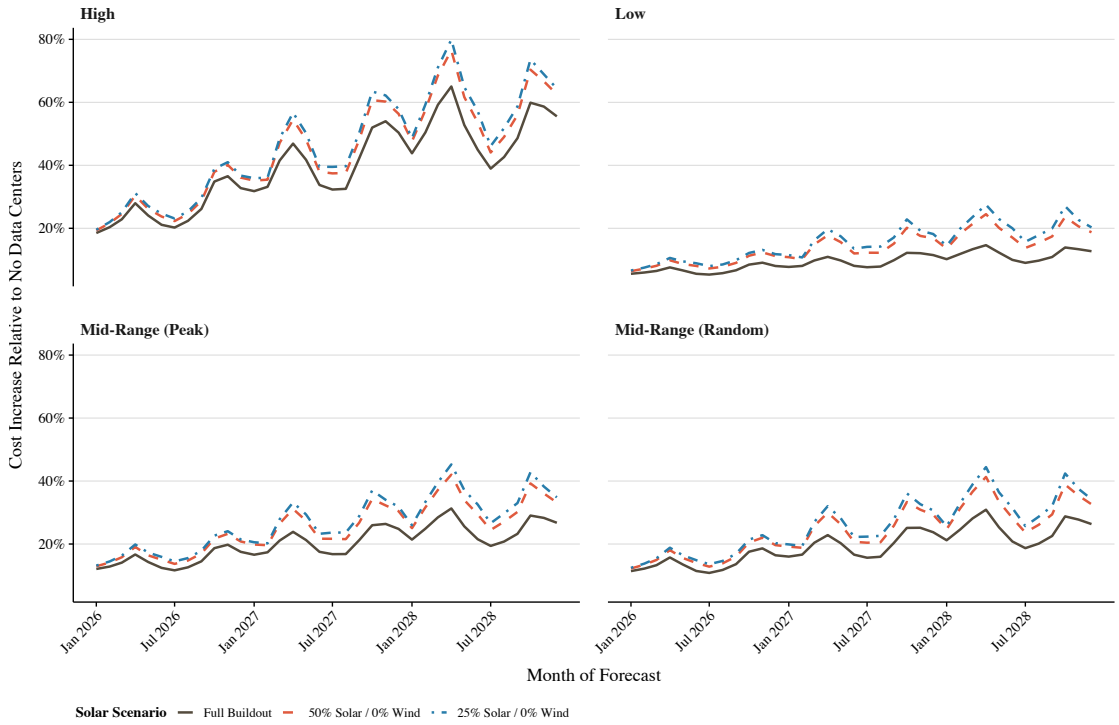


Notes: Graph shows the simulated monthly incremental grid emissions by data center utilization scenario.

Finally, we discuss how alternative assumptions about power plant entry and exit would impact our results. As data centers increase overall load, new power plants will be need to meet the increase in demand. The degree to which new generation assets connect to the grid can be important for determining wholesale market effects.

We first estimate the increase in thermal generation costs attributed to data centers if fewer sources of new solar and wind generation than expected connect to the grid. Figure 11 shows the effects of data center demand under two counterfactual supply scenarios. The black lines reproduce the baseline results shown in Figure 8. The dashed lines show the case where no additional wind generation is built and only 50% (red dashed) or 25% (blue dashed) of planned solar generation is built and available.

Figure 11: Projected Percentage Increase in Generation Costs
Under Alternative Renewable Build-Out, 2026–2028



Notes: Graph shows the simulated load-weighted average monthly incremental thermal generation costs by data center utilization scenario and renewable build-out scenario.

These alternative counterfactuals show that a slower build-out of wind and solar capacity would substantially amplify the market effects of data center expansion, even though data center demand is held fixed across scenarios. With fewer additional low-marginal-cost renewable resources available, incremental load is more likely to be served by higher-marginal-cost thermal generation. Given the convexity of the short-run electricity supply curve, this shift pushes dispatch into a steeper portion of the supply stack. Consequently, the same incremental increase in demand generates larger increases in generation costs and wholesale prices when renewable build-out falls short of expectations.

One important limitation to our modeling approach is that we treat entry and exit decisions of power plants as exogenous. Underlying this decision is an assumption that power plants that we observe as coming online in the next few years would do so even if there was no demand from data centers. We believe this assumption is reasonable in the short run. It takes several years to build

new power plants and connect them to the grid, a problem that has grown worse in recent years as interconnection times have increased significantly. In contrast, expected demand growth from data centers has only taken off very recently, implying that most of the projects in the interconnection queue were likely underdevelopment before data center demand was projected to be as large as current projections.

However, in the long run, it is plausible that new supply would enter specifically to address data center demand. Including these new entrants that are marginal to data center demand in the supply forecast for the 'no data center' counterfactual will artificially lower prices in this counterfactual and overstate the price and generation cost effects of data centers. To address this concern, we consider a case where 25% of new supply is marginal to datacenter demand and would not have entered had there been no data center demand.[7]

Appendix Figures A5 and A6 show how allowing for this endogenous supply build-out changes the baseline price and emissions effects of data centers. Price effects are largely unimpacted through early 2028, but begin to diverge in late 2028 as our main result overstates the effect of data centers by around 10% relative to the endogenous supply case. For emissions, the difference starts in 2027 as much of the new sources of supply that are slated to come online are renewable or efficient natural gas power plants. Removing them from the supply stack leads to higher emissions in our no data center counterfactual and compresses the gap between the counterfactuals with and without data center demand. For both prices and emissions, the effects of data centers are still meaningful. We also believe that treating a full 25% of new supply as marginal to datacenter demand represents a conservative estimate of the exogenous supply assumption.

A final concern is that our exogenous supply forecast includes prescribed power plant retirements, largely from coal plants. If prices increase as the modeling would suggest, the operational life of the assets may be extended in order to continue to make large profits. Additionally, there has been political pressures to keep these power plants open to meet rising demand. In Appendix Figures A7 and A8 we update our supply forecast to reverse planned retirements. Appendix Fig-

[7] We implement this by derating the capacity of all new entrants by 25% only in the 'no data center' counterfactual.

ure A7 shows the difference in price effects across the four data center demand scenarios. Prices are significantly lower in 2028 with no coal plant retirements. This price effect is concentrated in the small subset of markets that have substantial coal retirements, particularly MROW and NYUP. This however is offset by a further worsening of emissions effects as shown in Appendix Figure A8.

6 Discussion

We use our model to speak to several active policy conversations surrounding the rapid expansion of data centers. These concerns have emerged in a period characterized by rising electricity prices, an aging grid, and substantial uncertainty about the pace and location of new investment. We first discuss the total effect of data centers on retail rates and how the price increases relate to inflation. Data center connections will require large investments in transmission and distribution upgrades, however an outsized increase in demand could lead average fixed costs to decrease, leaving the effect on rates ambiguous. We next use our model to consider how market integration and data center flexibility can mitigate some of the increases in wholesale market prices.

6.1 Fixed Cost Allocation and Impact on Customer Bills

Our empirical analysis focused on the impacts to variable generation costs, but the electricity sector also requires large fixed expenditures to build, maintain, and expand generation, transmission, and distribution infrastructure. How these fixed costs are allocated across customers can differ across market designs and has important consequences for customer bills and the efficiency implications of electricity policies (Borenstein and Bushnell 2022; Borenstein et al. 2021). In vertically integrated systems, utilities generally recover both fixed and variable costs through retail rates, subject to regulatory review. In restructured markets, wholesale energy costs are determined in the market, while fixed transmission and distribution costs are recovered through regulated charges that are allocated across customer classes.

The effect of new data center load on the fixed-cost component of customer bills is theoretically ambiguous. Under simple average-cost allocation, total fixed costs are divided by the total kWh of electricity sold and recovered through retail rates. An increase in electricity consumption with no corresponding increase in fixed expenditures reduces the average fixed cost per unit of electricity sold. In this sense, additional load could lower the burden borne by each existing customer.[8] This outcome is especially relevant if data centers locate where there is pre-existing spare transmission and distribution capacity or flexibly decrease their consumption during periods of peak load (Norris et al. 2025).

At the same time, the scale, concentration, and clustering of proposed data center projects is likely to require substantial new expenditures on interconnection, transmission, substations, and local distribution networks. In that case, the numerator in the average-cost calculation rises as well, and potentially by more than the increase in energy sales. Whether data center growth lowers or raises the fixed-cost component of retail bills therefore depends on how strongly new load increases capital utilization relative to how much new infrastructure it necessitates.

Even if new data centers covered all the additional fixed costs associated with connecting to the grid, as many have promised to do, the downward pressure on fixed costs would need to offset the upward pressure from higher generation costs/prices to lower customer bills. Appendix E shows that this can only happen if the share of customer bills coming from fixed costs is sufficiently high. Specifically, even if new data center load does not increase fixed costs at all, it can only lower customer bills if

$$s \geq \frac{(1+d)}{1+d+\eta} \tag{5}$$

where s is the share of customer bills coming from fixed costs, d is the percent increase in electricity consumption from data centers, and η is the supply elasticity of electricity (percent increase in electricity supplied from a 1% increase in wholesale price).[9]

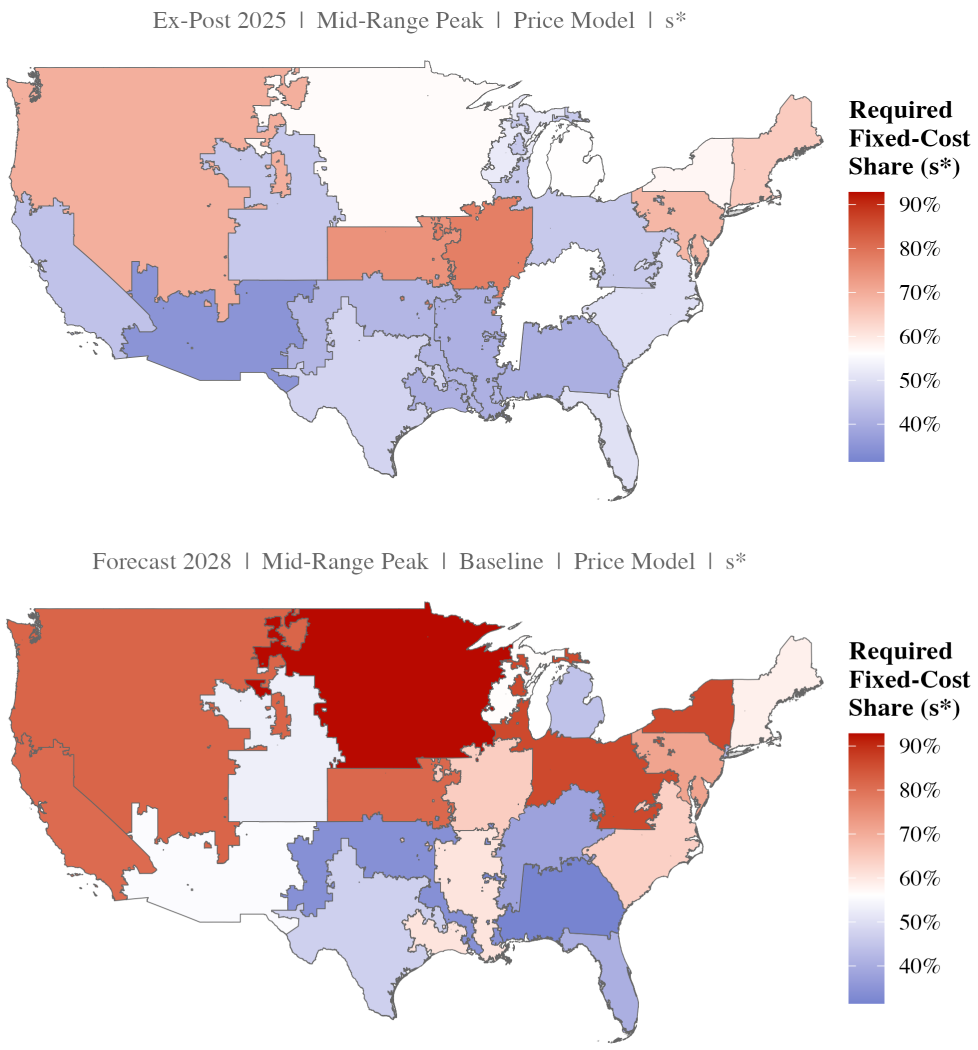
[8] Related concerns arise in the opposite direction when utilities face a shrinking customer base or declining sales, which can increase fixed cost recovery per customer; see, for example, Davis and Hausman (2022).

[9] This analysis also assumes that costs are recovered uniformly across all customer classes. In reality, commercial

Using the results from Sections 4 and 5, we calculate supply elasticities, demand increases, and therefore the threshold s such that Equation 5 holds with equality, s^*. Figure 12 maps these minimum fixed cost shares for each region in 2025 and 2028. For the median market, at least 56% of the customer bill would need to come from fixed costs for additional load to lower bills, even assuming that new load brings no additional fixed costs. There is variability in this figure across regions and over time, with many markets requiring fixed cost shares to be above 80%.

and industrial customers often face lower electricity rates than residential customers. In the limit, where datacenter are in a rate category for C&I customers that directly pays wholesale prices, there will be no fixed cost spreading from additional data center demand.

Figure 12: Threshold Fixed Cost Share Required for Rate Relief



Notes: This map shows the minimum fixed cost share such that it would be possible for new load to lower residential bills, assuming that new load does not increase total fixed costs at all. The analysis uses wholesale prices/system lambda as the relevant measure of variable cost pass through to end-use customers. This assumption is reasonable for parts of the country with wholesale electricity markets. In regions that did not restructure electricity markets, average variable costs may be a more appropriate measure.

This tradeoff between fixed cost spreading and variable cost increases is central to current policy debates. In some jurisdictions, legislators and regulators have proposed mechanisms to shield existing customers from network upgrade costs associated with large-load interconnections. At the same time, several data center developers have indicated a willingness, at least in principle, to fund portions of network upgrades required to serve their facilities. This analysis shows that these measures on their own may not be enough. If variable costs comprise a significant fraction

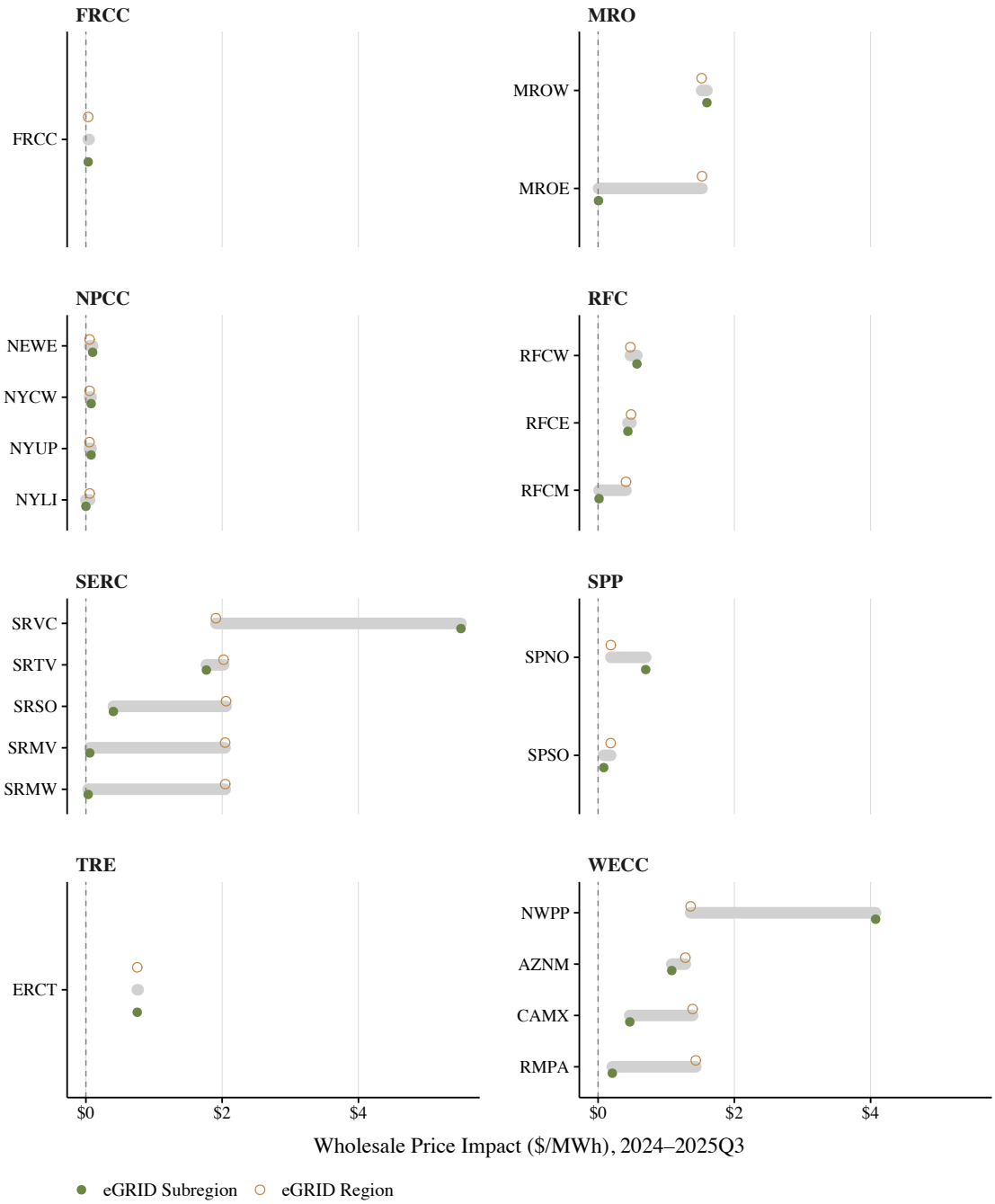
of retail bills, it may be necessary for data centers to make additional transfer payments in order to prevent rates from increasing for existing customers.

6.2 Transmission Constraints and the Value of Market Integration

We can also use our results to speak to the role of transmission and market integration in mitigating localized price pressure and accommodating load growth. Load growth is highly uneven across regions and low-cost generation and spare capacity are often located elsewhere. This creates a natural role for transmission: moving relatively cheap electricity from lower-demand areas into the most constrained locations. Achieving this in practice, however, requires substantial interstate and possibly interregional transmission expansion, which is costly, slow to permit, and subject to regulatory and political constraints.

To quantify the potential importance of these frictions, Figure 13 presents a counterfactual in which markets are seamlessly integrated and transmission congestion is eliminated. Operationally, this means that the hourly least-cost dispatch problem is solved at the level of the aggregated eGRID region instead of subregion, allowing the model to dispatch the lowest-cost available generation across a wider geographic footprint. In the regions that are most exposed to data-center-driven scarcity in the baseline, such as SRVC, this integration substantially attenuates the local price effects of data center demand. At the same time, the gains are not uniform. Despite a decrease in the nationwide average price, some regions experience higher prices under aggregation because previously idle, out-of-merit units are now called upon more frequently to serve demand elsewhere in the larger integrated market. In that sense, market integration does not eliminate scarcity; rather, it redistributes it across space and reduces the extent to which scarcity is concentrated in a small number of constrained subregions.

Figure 13: Change in Price Effects with Market Aggregation



Notes: This figure shows the change in the simulated load-weighted average wholesale market price impact by market aggregation level and subregion. Effects for the mid-peak scenario are shown.

We do not attempt to quantify the capital costs, siting barriers, or political economy constraints associated with the transmission build-out needed to support this level of integration. Nonetheless, the exercise highlights that transmission constraints are first-order for the incidence of data-center-

driven price increases.

This exercise also serves as a robustness check on our assumptions rergarding interregional transmission. Although eGRID subregions are defined in order to minimize flows between bordering regions, some trade still occurs. When we re-dispatch generation under a no-datacenter counterfactual, we assume that local generation is marginal. As a result we might over-state cost/price impacts in subregions that are marginal importers and under-state impacts in marginal exporting subregions. Figure 13 displays the directions in which our results may be biased, and provides bounds on these impacts. Because we allow for uncongested trade within the integrated regions, it is reasonable to expect the size of the bias to be smaller than the difference in impacts shown above. Thus, this exercise not only highlights the role of market integration in mitigating localized pressures, but also affirms the meaningful impacts already seen in areas with concentrated data center load.

6.3 Geographic Compute Flexibility as Virtual Transmission

Physical transmission is not the only potential margin of adjustment. A conceptually related alternative is greater flexibility in where compute is performed. If computational workloads can be curtailed, deferred, or spatially reallocated across geographically dispersed data centers, then compute flexibility can act, at the margin, like a form of virtual transmission. This shifts effective electricity demand away from stressed regions and toward regions with lower prices and more spare capacity.

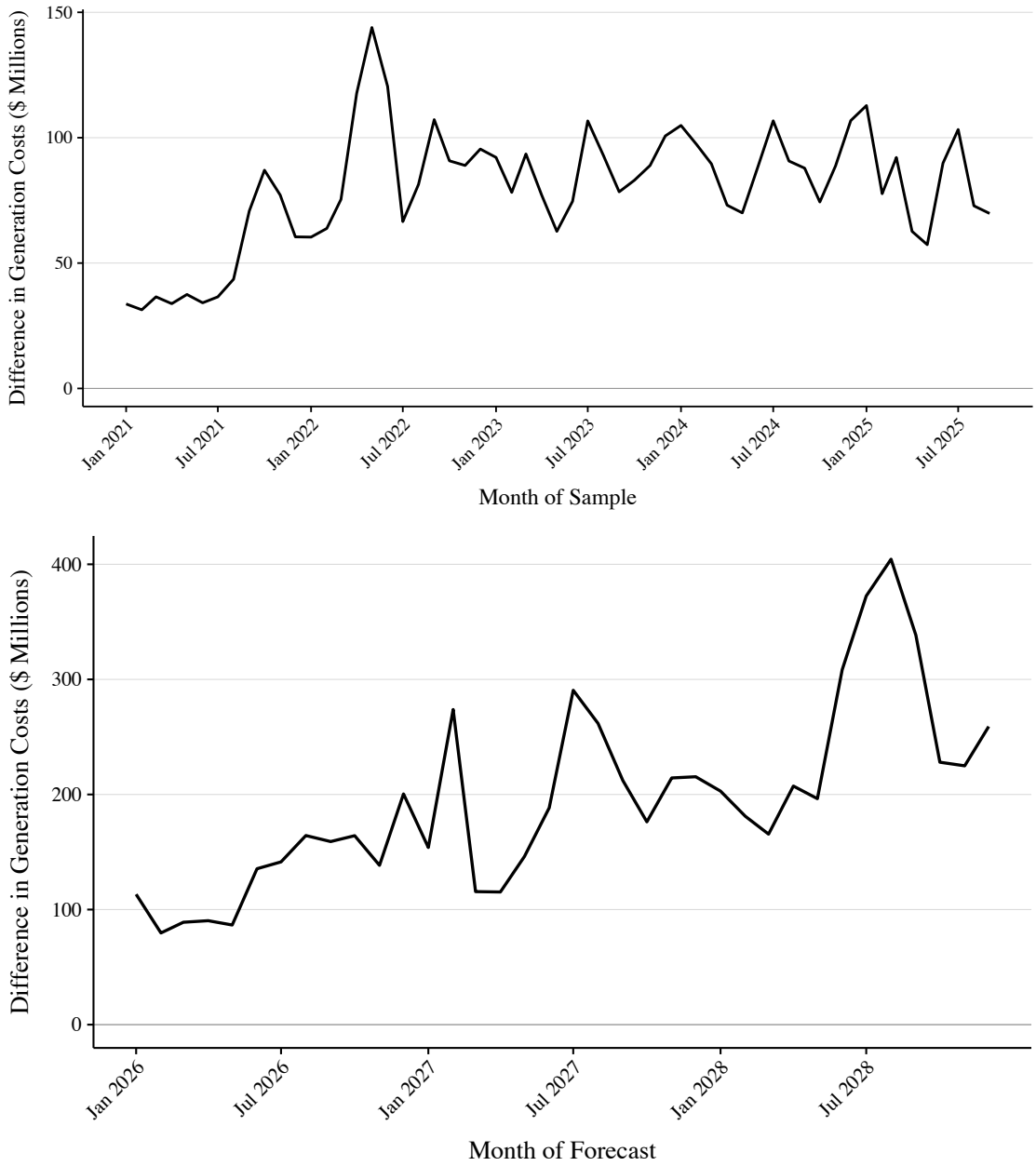
In principle, increased flexibility can improve grid reliability and lower wholesale prices, but it requires sufficient excess and geographically diverse compute capacity, network capability, and tolerance for latency and other operational constraints. In practice, many data center operators market their services on the basis of extremely high reliability, making frequent and unplanned curtailment costly. In this sense, spatial compute flexibility may substitute for transmission only if firms are willing to maintain some level of costly overbuild of data center capacity across multiple regions or institute some prioritization of tasks.

To illustrate the potential magnitude of this margin, we compare observed data center siting against the geographic allocation of compute chosen by a social planner minimizing short-run electricity operating costs.[10] This counterfactual abstracts from latency, construction costs, labor availability, interconnection constraints, and other factors that shape actual siting decisions. We solve a national dispatch problem without regional transmission constraints using excess thermal capacity that remains idle in the no-data-center counterfactual and allocate total hourly data center demand to the lowest-cost available generation nationwide. The resulting hourly generation pattern identifies where compute would be served if short-run electricity operating cost were the sole objective.

Figure 14 shows the magnitude of potential generation cost savings from allowing data center demand to be geographically flexible. From 2021 through 2025Q3, the monthly savings from the cost-minimizing spatial allocation mostly fluctuated between $50 and $100 million a month. Looking forwards, these potential costs savings increase as overall data center demand increases.

[10]For full details on how we solve the social planner's problem, see Appendix F.

Figure 14: Difference in Generation Costs Between Observed and Spatially Optimal Compute



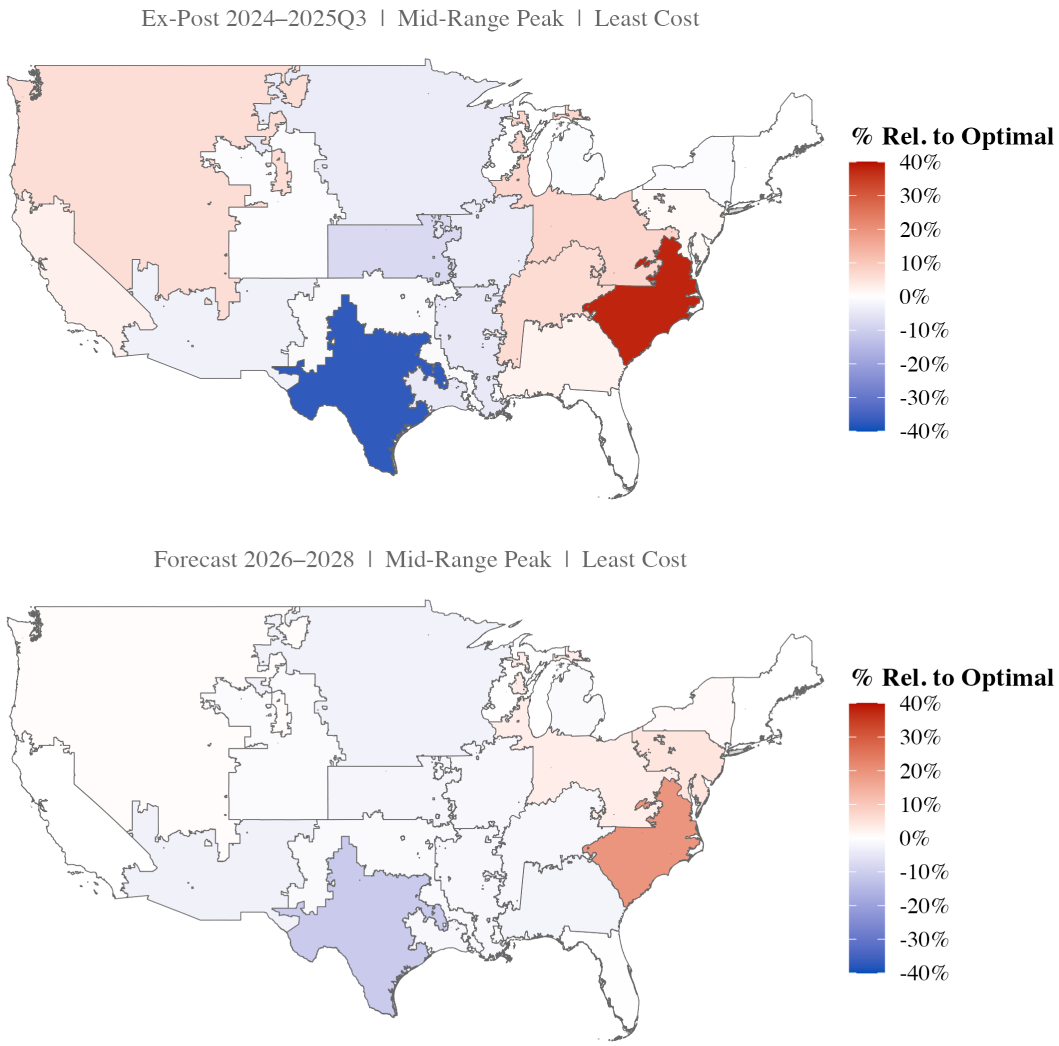
Notes: These charts show the difference in monthly electricity generation costs between observed data center compute under the mid-peak utilization scenario and the cost-minimizing spatial allocation of compute which allows for data center demand to flexibly move to the the lowest cost generation sources available.

Figure 15 shows that the difference in geographic allocation between observed and the optimal allocation in our mid-peak demand scenario is large. Over 2024–2025Q3, ERCOT accounts for 11.7 percent of realized data center compute in the data, whereas the social planner would allocate approximately 49 percent there—a gap of roughly 37 percentage points. By contrast, SRVC

supplies 38.8 percent of realized compute over this period, while the cost-minimizing allocation assigns it only 0.78 percent due to the high pre-existing wholesale prices. Under the planner allocation, just four regions—ERCT, MROW, AZNM, and SPNO—would serve more than 75 percent of compute demand. These wedges imply that non-energy siting constraints are first-order determinants of realized data center geography and, in turn, of the regional distribution of wholesale price pressure.

The same qualitative pattern persists in the forecast period, although the allocation becomes somewhat less concentrated. For example, SRVC accounts for 19.36 percent of installed compute in the forecast but only 3.4 percent under the least-cost allocation, while ERCOT rises from 21.8 percent installed to 30.7 percent under the benchmark. Under the forecasted system, 75 percent of compute is served by eight regions—ERCT, RFCW, SRTV, AZNM, NWPP, MROW, SRSO, and SRVC—rather than four. This change suggests that future capacity expansion and evolving system conditions may reduce, but do not eliminate, the degree to which realized data center geography diverges from a short-run cost-minimizing allocation.

Figure 15: Distribution of Observed Compute Compared to the Optimal Allocation



Notes: Maps show the percentage point difference in the compute market share by subregions between the optimal spatial allocation and the observed distribution.

7 Conclusion

This paper provides one of the first empirical assessments of how expanding data center load affects wholesale electricity market outcomes in the United States. Using an hourly least-cost dispatch model built from unit-level data covering the continental U.S., we estimate how data center load changes generation costs, competitive benchmark wholesale prices, and carbon emissions in both the recent historical period and a set of forward-looking counterfactuals.

Three main findings emerge. First, data center expansion has already increased generation costs, competitive wholesale benchmark prices, and emissions in several regions. We find that generation costs have increased by 5–15% from 2021-2025, depending on data center utilization assumptions. These effects are not uniform across markets and are largest in subregions where data center growth coincides with relatively tight thermal capacity, such as Virginia and the Carolinas. Second, the impacts of data center load are nonlinear. Because of highly convex short-run supply curves, even moderate increases in load can generate disproportionately large increases in marginal generation costs when the system is operating near scarcity. These effects are most pronounced in seasons when outages or maintenance reduce available capacity. Third, the forward-looking results show that the data center build-out will have increasing effects on wholesale markets. Uncertainty in how much proposed data center capacity will actually be built is a main source of uncertainty, but in all scenarios tested, the model implies meaningful increases in wholesale market pressure.

The broader policy implication of our model is that the economic consequences of data center expansion depend not only on the size of the increase in demand, but also on where it is located, when it operates, and how quickly new supply can be built. We contribute to active policy discussions by using our model to show how transmission expansion and interconnection policy, resulting in broader market integration, can lead to lower nationwide effects, albeit at the expense of consumers in previously shielded subregions. We also highlight that data center flexibility can play a role in reducing the effects by comparing the observed build-out to a optimal spatial allocation of compute that prioritizes least-cost computing.

References

Benetton, Matteo, Giovanni Compiani, and Adair Morse, "When Technology Processors Come to Town: High Electricity Use Spillovers to the Local Economy," Working Paper 31312, National Bureau of Economic Research June 2023. NBER Working Paper No. 31312.

Bogmans, Christian, Patricia Gomez-Gonzalez, Ganchimeg Ganpurev, Giovanni Melina, Andrea Pescatori, and Sneha Thube, "Power Hungry: How AI Will Drive Energy Demand," IMF Working Paper 2025/081, International Monetary Fund April 2025.

Borenstein, Severin, "The Trouble With Electricity Markets: Understanding California's Restructuring Disaster," *Journal of Economic Perspectives*, February 2002, *16* (1), 191–211.

_ , "What Will Data Centers Do To Your Electric Bill?," Energy Institute Blog, UC Berkeley September 2025.

_ **and James B. Bushnell**, "Do Two Electricity Pricing Wrongs Make a Right? Cost Recovery, Externalities, and Efficiency," *American Economic Journal: Economic Policy*, November 2022, *14* (4), 80–110.

_ **, James B Bushnell, and Frank A Wolak**, "Measuring market inefficiencies in California's restructured wholesale electricity market," *American Economic Review*, 2002, *92* (5), 1376–1405.

_ **, Meredith Fowlie, and James Sallee**, "Designing Electricity Rates for An Equitable Energy Transition," Energy Institute at Haas Working Paper WP 315 February 2021.

Bushnell, James, "A mixed complementarity model of hydrothermal electricity competition in the western United States," *Operations research*, 2003, *51* (1), 80–93.

Davis, Lucas W and Catherine Hausman, "Who Will Pay for Legacy Utility Costs?," *Journal of the Association of Environmental and Resource Economists*, August 2022, *9* (6), 1047–1085.

Feher, Adam, Emilia Garcia-Appendini, and Roxana Mihet, "Is AI Trained on Public Money? Evidence from U.S. Data Centers," November 2025. Working paper.

Gargano, Antonio and Marco Giacoletti, "Subsidizing the Cloud: U.S. State Incentives to Data Centers," Working Paper 25-588, SSRN eLibrary 2025.

Green, Alastair, Humayun Tai, Jesse Noffsinger, Pankaj Sachdeva, Arjita Bhan, and Raman Sharma, "How data centers and the energy sector can sate AI's hunger for power," Technical Report, McKinsey & Company September 2024.

Ham, Dasom, Owen Kay, and Catherine Hausman, "Power Flows, Part 2: Transmission Lowers US Generation Costs, But Generator Incentives Are Not Aligned," Technical Report, Resources for the Future 2025.

Hausman, Catherine, "Power flows: Transmission lines, allocative efficiency, and corporate profits," *American Economic Review*, 2025, *115* (8), 2574–2615.

Knittel, Christopher R, Juan Ramon L Senga, and Shen Wang, "Flexible Data Centers and the Grid: Lower Costs, Higher Emissions?," Technical Report 2025-14, Massachusetts Institute of Technology Center for Energy and Environmental Policy Research July 2025.

Mamkhezri, Jamal, Xiaochen Sun, and Yuting Yang, "The Hidden Cost of the Cloud: Data Centers and Electricity Market Inefficiency," Working Paper 25-660, USAEE / IAEE Working Paper Series 2025.

Muller, Nicholas Z, "Measuring the Impact of Data Centers in the United States Economy: Monetary Damage from Air Pollution and Greenhouse Gas Emissions," *NBER Working Paper Series*, April 2026.

Norris, Tyler H., Tim Profeta, Dalia Patino-Echeverri, and Adam Cowie-Haskell, "Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power

Systems," Report NI R 25-01, Nicholas Institute for Energy, Environment & Sustainability, Duke University 2025.

Ross, Martin T. and Jackson Ewing, "Data Centers and Generation Capacity over the Next Decade: Potential Benefits of Flexibility," Technical Report NI 26-04, Nicholas Institute for Energy, Environment & Sustainability, Durham NC 2026.

Shehabi, Arman, Sarah J. Smith, Alex Hubbard, Alexander Newkirk, Nuoa Lei, Md AbuBakar Siddik, Billie Holecek, Jonathan G. Koomey, Eric R. Masanet, and Dale A. Sartor, "2024 United States Data Center Energy Usage Report," Technical Report, Lawrence Berkeley National Laboratory December 2024.

Wade, Cameron, Mike Blackhurst, Joe DeCarolis, Anderson de Queiroz, Jeremiah Johnson, and Paulina Jaramillo, "Electricity Grid Impacts of Rising Demand from Data Centers and Cryptocurrency Mining Operations," Report, Scott Institute for Energy Innovation, Carnegie Mellon University June 2025.

A Appendix: Tables

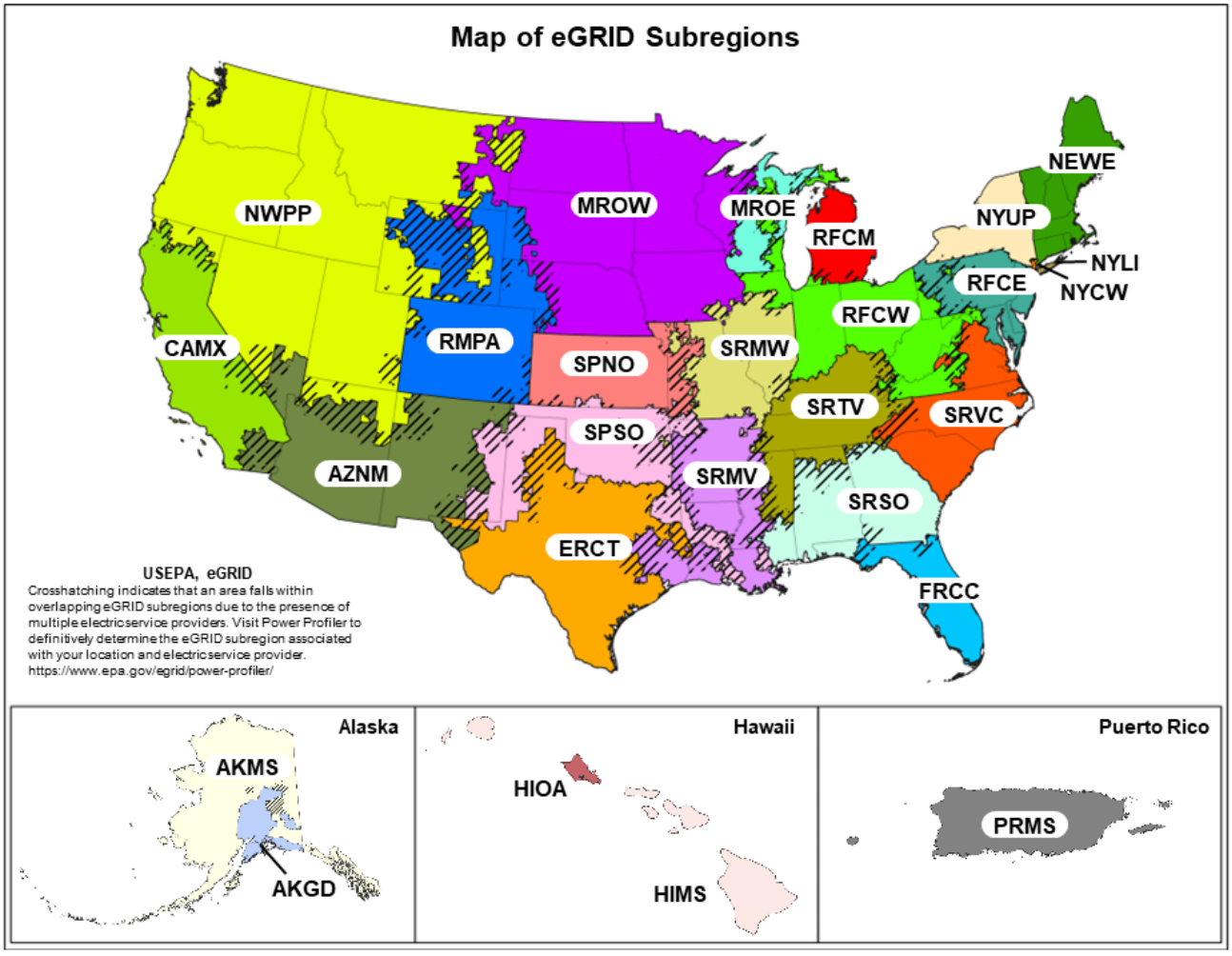
Appendix Table A1: County Characteristics: Data Center vs. Non-Data-Center Counties

	DC counties	Non-DC counties	Difference
Population	508,698	68,056	440,642***
Median household income (USD)	81,069	64,439	16,630***
Poverty rate (%)	12.06	14.48	-2.42***
Unemployment rate (%)	4.70	4.69	0.01
Bachelor's degree or more (%)	33.88	23.01	10.86***
Non-Hispanic White share (%)	62.41	75.26	-12.85***
Median age	38.38	42.01	-3.63***
Industrial electricity price (cents/kWh)	8.07	7.70	0.37*
N counties	271	2751	

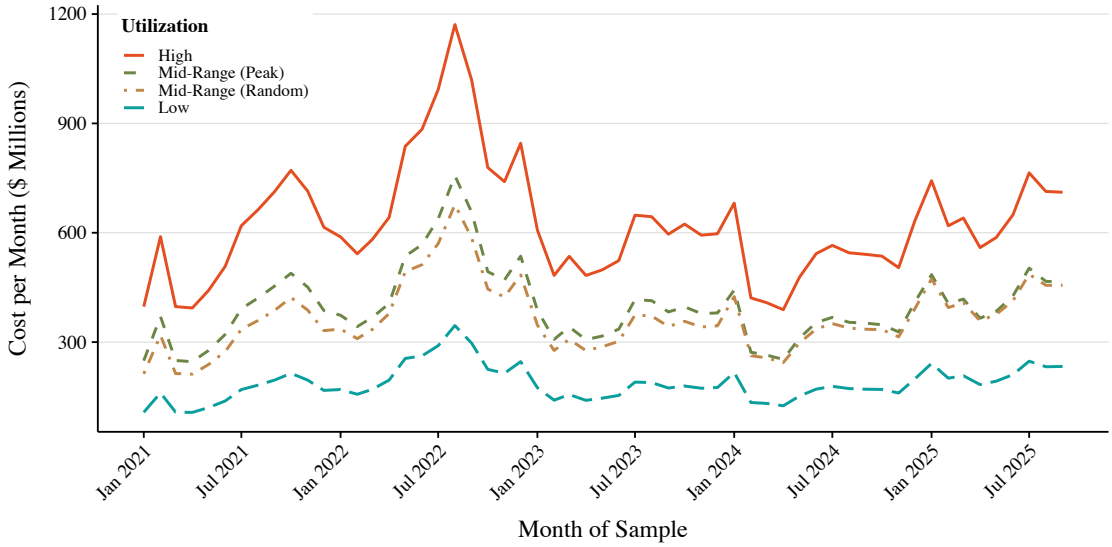
Notes: DC counties are counties with at least one non-cancelled data center in the Cleanview sample. Demographics are from ACS 2019–2023 5-year estimates. Industrial electricity price is from EIA-861 and is constructed at the county level using utility-state industrial revenue and sales mapped to counties via service territories, then averaged over 2019–2023. Difference reports mean(DC) - mean(Non-DC). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Appendix: Figures

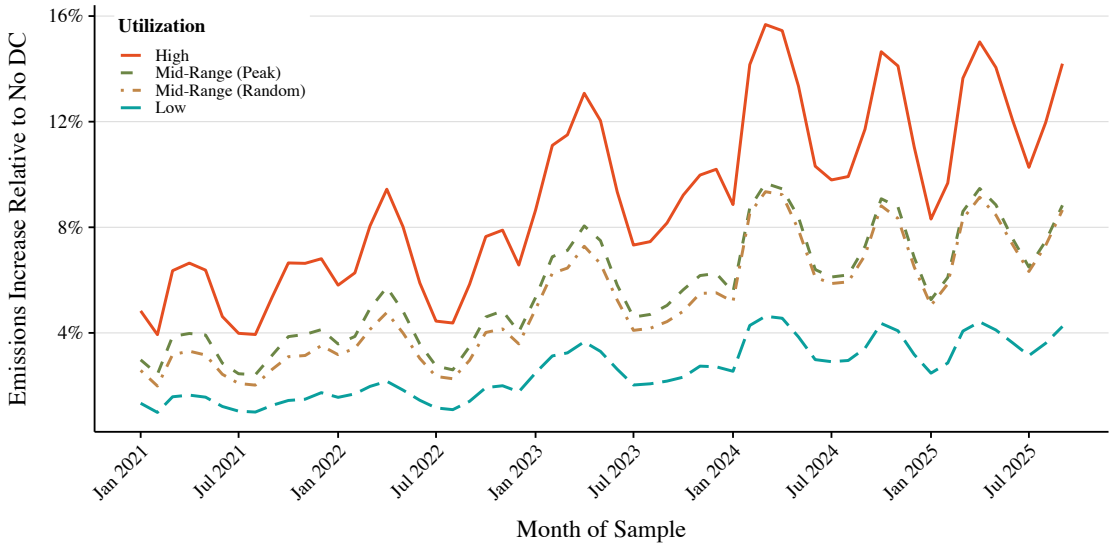
Appendix Figure A1



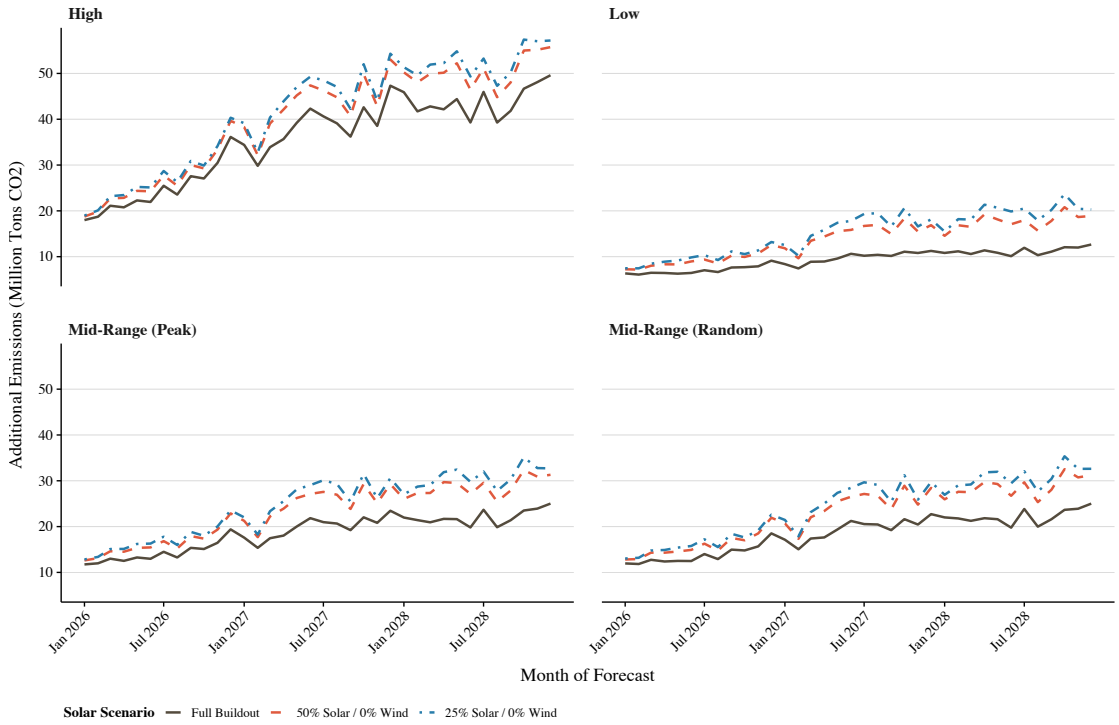
Appendix Figure A2: Monthly Incremental Costs, 2021–2025Q3



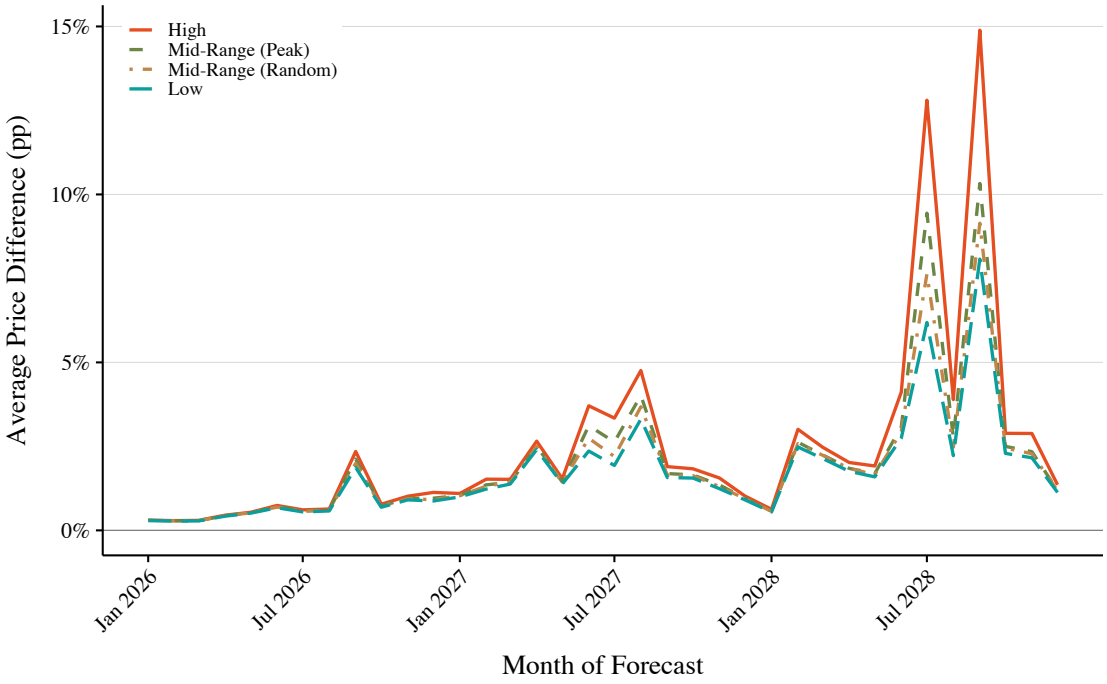
Appendix Figure A3: Monthly Percentage Increase in CO_2 Emissions, 2021–2025Q3



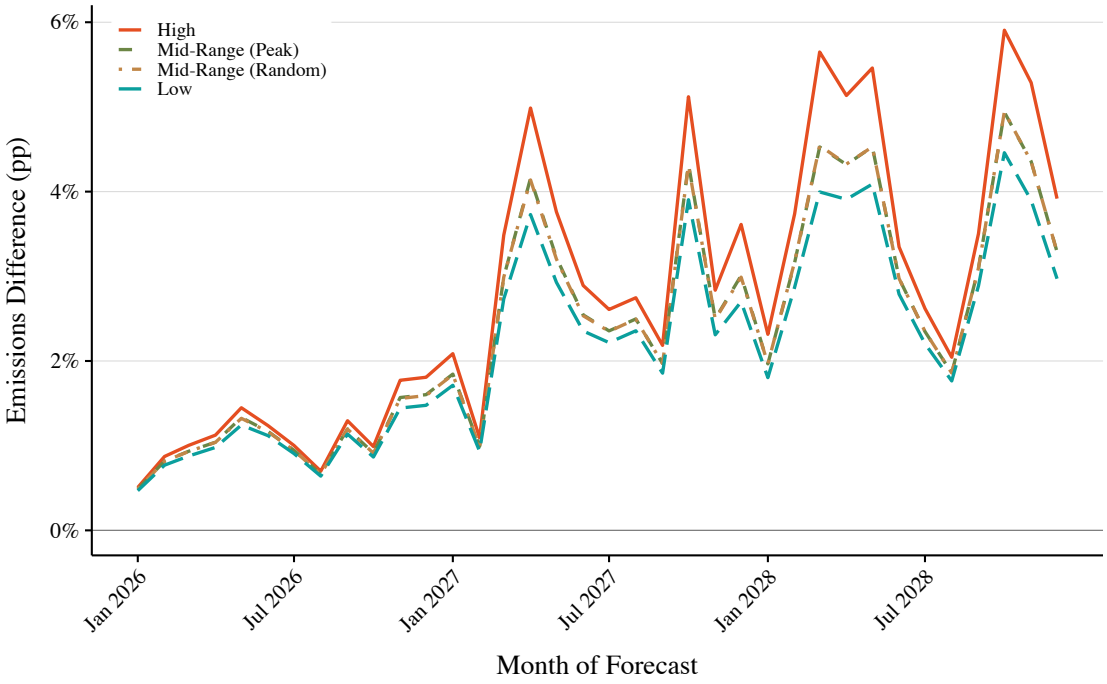
Appendix Figure A4: Forecasted Incremental Carbon Emissions under Alternative Renewables Buildout, 2026–2028



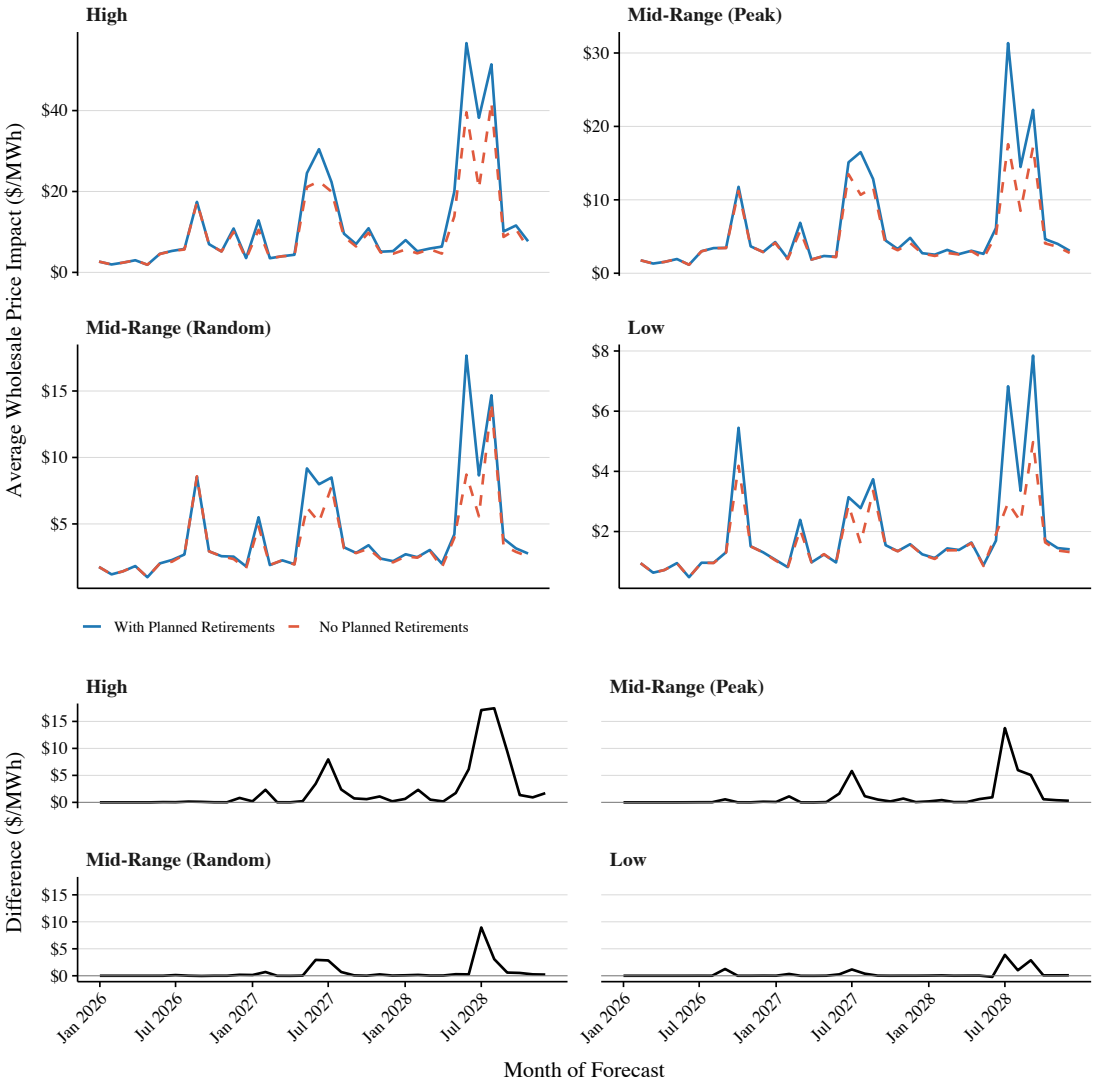
Appendix Figure A5: Difference in Forecasted Wholesale Market Price under Endogenous Build-out Assumption, 2026–2028



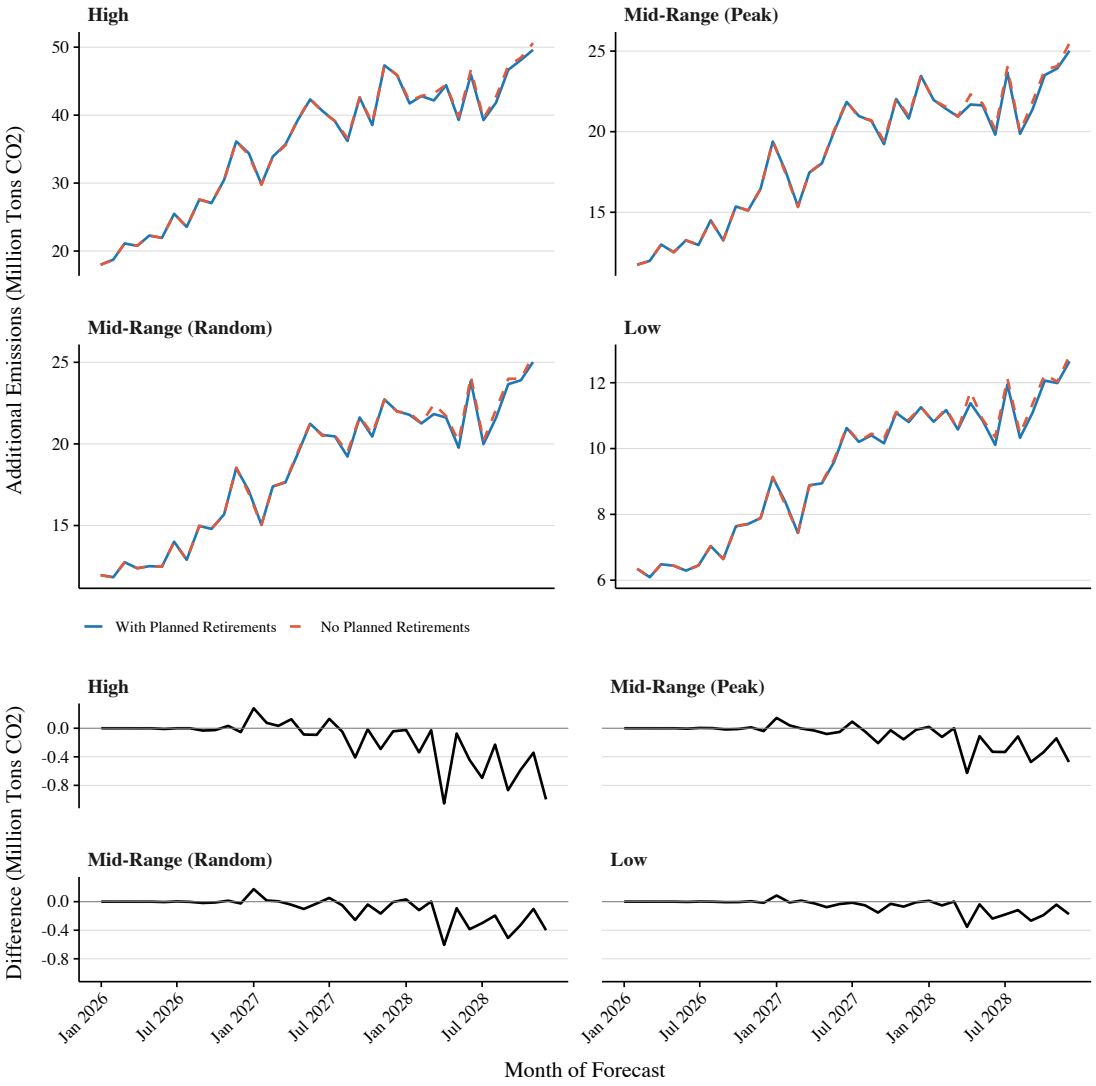
Appendix Figure A6: Difference in Forecasted Carbon Emissions under Endogenous Build-out Assumption, 2026–2028



Appendix Figure A7: Forecasted Wholesale Market Price under Alternative Coal Retirement, 2026–2028



Appendix Figure A8: Forecasted Incremental Carbon Emissions under Alternative Coal Retirement, 2026–2028



C Appendix: Ex Post Model Fit

Here, we discuss the goodness of fit of our baseline model, simulated on observed outcomes from 2021–2025.[11] We compare the results from our eGRID subregion model to those from the model using the aggregated regional definitions and show that, in general, the subregion model appears to match the observed data better.

C.1 Quantities

Table A2 shows that our model using eGRID subregions closely matches first and second moments of observed generation quantities. In addition, our model matches the reality that the majority of capacity is offline at any given moment, reflected in our median quantity being zero. Our model using the aggregate eGRID regions also performs well in this respect.

Appendix Table A2: Observed Versus Simulated Generation, Summary Statistics

		First and Second Moments		Percentiles		
	Count	Mean	SD	25th	Median	75th
Observed generation	120,076,459	73.78	141.35	0	0	88.49
Subregion simulation	120,076,459	73.76	154.84	0	0	31.88
Aggregate simulation	120,076,459	73.77	155.73	0	0	35.54

This table summarizes observed versus simulated generation at the unit × hour level. Sample covers 2021–2025.

Table A3 presents the correlations between observed and simulated generation at the unit × hour, unit × month, and unit × year levels. Despite the fact that we apply uniform derates and stochastic outages across all units, simulated output is still highly correlated with observed generation. Figure A9 provides a useful visualization for the mechanisms that cause our simulated generation to deviate from observed, by construction. The density of observations along the x-axis represent observations where we stochastically removed units from the supply curve while they

[11] Given the similarity of our model to those of Hausman (2025) and Ham et al. (2025), much of the discussion that follows closely mirrors the discussion of model fit in those papers.

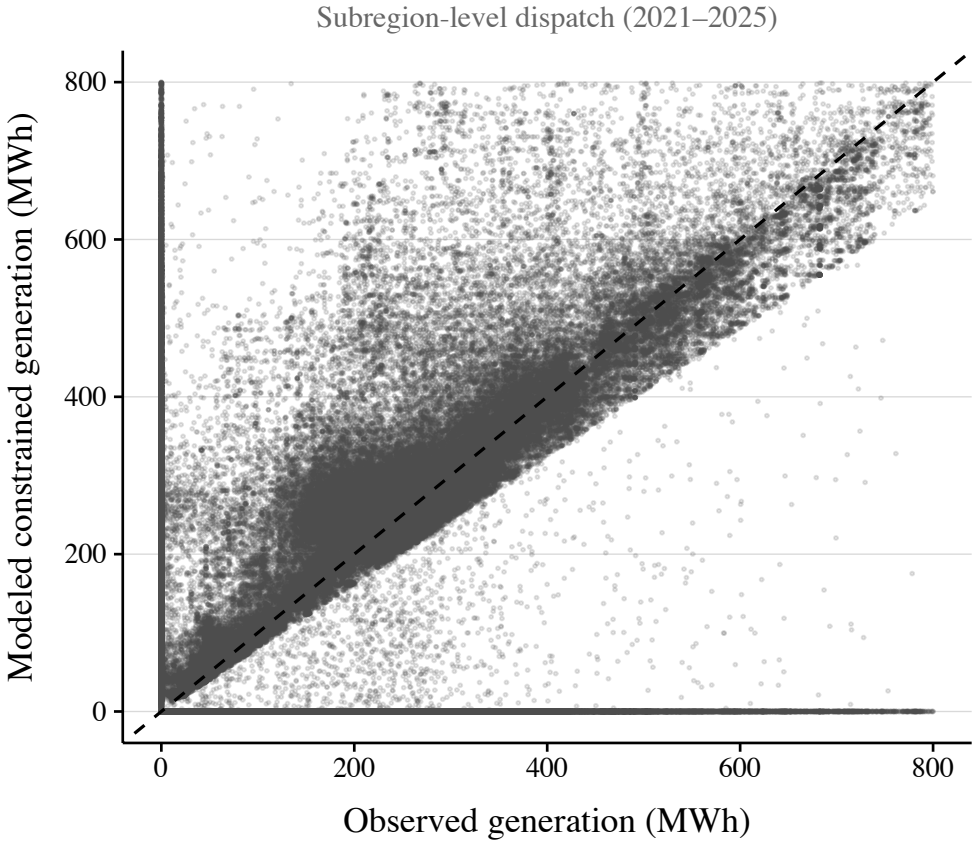
were actually on in reality. Similarly, the points along the supply curve are those observations where the unit was offline in reality, and our model left them available for dispatch. The set of off-diagonal points below the 45-degree line reflect the uniform derates we apply to each unit by fuel type. When we smooth over these structural mismatches by aggregating to the monthly and yearly levels, unit-level correlations rise substantially. These values are reported in the middle and right-most columns of Table A3. Unit-level correlations in our subregion model are slightly higher than the aggregated model, which suggests that the boundaries of the eGRID subregions delineate more electrically-consistent areas.

Appendix Table A3: Correlation Between Observed and Simulated Generation: Aggregate vs. Subregion

	Correlation with Observed Generation		
Version	Hourly	Monthly (unit × month)	Yearly (unit × year)
Subregion	0.631	0.787	0.863
Aggregate	0.617	0.772	0.848

Hourly correlations are computed at the unit × hour level. Monthly and yearly correlations average over hours within unit × month and unit × year, respectively. Sample covers 2021–2025.

Appendix Figure A9: Model Fit: Hourly Observed Versus Simulated Output



Note: 0.5% random subsample of hour-by-unit observations.

C.2 Competitive Wholesale Prices

We are also able to compare our competitive wholesale price to the set of restructured markets in our setting: CAISO, ERCOT, ISONE, MISO, NYISO, PJM, and SPP. All of these markets report some version of an unconstrained price of energy —the component of each node's locational price that excludes congestion costs and line losses. Thus, it represents the shadow price of the system-wide power balance constraint, just as the competitive wholesale price in our simulations.[12] The actual system energy price is derived from generator bids, so in some hours, especially when local constraints bind, our model may diverge from the observed prices. However, in the majority of hours, when there are limited constraints on the system, we should expect system prices to reflect

[12]Because the optimization problem that each market solves is more complex than our simple LP, how they calculate this value tends varies in practice.

the marginal cost of the generator on the margin.

In Table A4 we compare how our simulated prices match observed moments. Panels A and B pool all markets together while Panel C compares the moments of individual wholesale markets. We compute average prices for each wholesale market using the hourly simulated price of each subregion that overlaps the wholesale market, weighted by its share of that market's load.[13] We weight observed prices by total load in that market. Our choice to simulate a static model with constant marginal cost up to each unit's capacity with no transmission constraints within an eGRID subregion means we do not capture the intertemporal and spatial dynamics that can cause system prices to vary substantially. Panels B and C exclude observations from hours when prices fall in the top and bottom 1% of the observed distribution.

Panel B shows that for the vast majority of hours the simulated prices replicate observed moments well, at a national level. While the average simulated price lies just $1.33 below the observed price, just a 3% deviation. At the market level, the simulated moments largely track the observed moments, particularly in regions featured prominently in this paper. Namely, CAISO, ERCOT, MISO, PJM, and SPP all closely match the observed moments after filtering out extreme observations. While the mean prices in ISONE and NYISO diverge, the median prices, particularly in ISONE, are still close. Given the simulated in both markets falls below the observed mean, our results should tend to understate the effect of data centers on wholesale prices in these regions, given the convex nature of the market supply curve.

Hourly price correlations follow a similar pattern to the quantity discussion, but highlight additional modelling assumptions. With no intra-regional transmission capacity constraints or generator constraints such as start-up costs, ramp rates, or minimum uptimes, the simulated prices will follow a much flatter pattern than is observed in reality. Aggregating to the monthly-of-sample and annual levels leads to much higher correlations between simulated and observed prices. The analysis in this paper primarily concerns itself with the monthly and annual levels of analysis —levels, at which, simulated prices are highly correlated with observed prices.

[13] This approach is similar in spirit to how many ISOs/RTOs compute their system-level energy price. It is often a load weighted average of several load zones.

Appendix Table A4: Observed Versus Modeled Wholesale Electricity Prices

	N	Mean	SD	25th pct.	Median	75th pct.
A: All Observations						
Observed price	269,536	45.43	87.35	22.04	32.36	51.53
Modeled price	269,536	39.99	19.39	28.93	34.88	44.58
B: Excl. Top/Bottom 1%						
Observed price	259,712	40.31	30.45	22.32	32.40	50.70
Modeled price	259,712	38.98	14.87	29.08	34.90	44.23
C: By ISO/RTO, Excl. Top/Bottom 1%						
CAISO — observed	36,754	48.17	32.77	29.66	43.28	59.49
CAISO — modeled	36,754	48.77	14.50	36.55	45.88	55.17
ERCOT — observed	39,044	36.09	30.17	18.87	26.30	40.81
ERCOT — modeled	39,044	35.81	17.22	24.88	29.10	40.63
ISONE — observed	39,584	52.05	38.81	25.93	38.05	59.80
ISONE — modeled	39,584	43.50	12.70	34.76	38.91	47.51
MISO — observed	25,363	32.75	18.99	21.48	26.44	35.26
MISO — modeled	25,363	31.78	6.83	27.75	30.03	33.08
NYISO — observed	39,569	58.86	50.18	27.30	39.42	64.82
NYISO — modeled	39,569	43.16	18.50	30.42	36.26	47.38
PJM — observed	39,713	42.25	26.45	23.24	32.66	49.55
PJM — modeled	39,713	38.81	10.66	32.40	35.27	39.86
SPP — observed	39,685	31.02	26.83	15.68	24.35	37.87
SPP — modeled	39,685	33.02	13.09	24.77	27.95	34.78

This table reports moments of observed market prices and modeled competitive wholesale prices, weighted by observed hourly load (MWh). Observed prices use the energy component of each ISO/RTO's LMPs from their real-time price data for 2021–2025Q3. Modeled prices are the shadow price of the power-balance constraint from the least-cost dispatch model. All prices in $/MWh. Panel A includes all hours for which at least one price is available. Panel B removes prices above and below the 99^{th} and 1^{st} percentiles, respectively. Panel C disaggregates Panel B by ISO/RTO. No data for 2021 MISO prices.

D Appendix: Forward Looking Model

In order to solve for market prices in future pieces, we must construct hourly demand and supply forecasts out through 2028. We detail their construction below.

Appendix Table A5: Correlations Between Hourly Observed And Modeled Wholesale Prices

	Hourly	Monthly	Annual
A: All Observations			
ISO/RTO Price × hour	0.305	0.517	0.677
B: Excl. Top/Bottom 1%			
ISO/RTO Price × hour	0.504	0.629	0.773
C: By ISO/RTO, Excl. Top/Bottom 1%			
CAISO	0.478	0.496	0.717
ERCOT	0.470	0.693	0.710
ISONE	0.393	0.446	0.688
MISO	0.481	0.738	0.881
NYISO	0.425	0.574	0.861
PJM	0.684	0.854	0.898
SPP	0.531	0.783	0.761

This table reports the Pearson correlation between the observed energy component of the LMP and the modeled competitive wholesale price at the hourly level. Hourly correlations use the full panel of ISO/RTO × hour observations. Monthly and annual correlations first average prices within ISO/RTO × hour × month-of-sample and ISO/RTO × hour × year, weighted by observed load. Panel A uses all available observations. Panel B removes prices above and below the 99^{th} and 1^{st} percentiles, respectively. Panel C disaggregates Panel B by ISO/RTO. No data for 2021 MISO prices.

D.1 Forward Looking Demand Scenarios

To calculate a demand forecast through 2028, we combine data on current hourly production from EPA eGRID and EIA-930, and annual electricity consumption from the EIA Annual Energy Outlook (AEO).

Our goal is to forecast electricity demand from all sources except data centers. We start with the 2021 observed annual net generation (TWh) by eGRID region as reported in the 2021 EPA eGRID model. We then subtract off annual data center demand in 2021 using the observed 2021 data center capacity from cleanview and market wide utilization rates reported from Bloomberg NEF/DC Byte. We use annual forecasts of electricity demand by electricity market module (EMM)

region from the 2021 EIA AEO to calculate year-over-year growth rates from 2021 through 2028. As these forecasts were made prior to the massive performance improvements in AI and subsequent build out of data centers, we use the 2021 numbers as representing expected electricity growth absent modern large-scale data centers but including expectations of growth in demand from electrification and the energy transition.[14] Year-over-year growth rates are calculated for the EMM regions reported in the AEO and then mapped to EPA eGRID regions.[15] We then apply these growth rates to the 2021 net generation without data center demand calculated above to estimate annual demand for years 2026-2028 *net* of data center demand.

To apportion this annual demand number across hours within a year we use data from the EIA-930 form which reports hourly net generation by balancing authority for January 2021 through 2025. For each balancing authority, we generate a share for each hour of the year which is the percent of annual net generation consumed in that specific hour. Using these observed shapes, we then construct a panel of forecast load shapes for our forward-looking years by performing a block bootstrap at the month level to allow for temporal correlation in demand. This results in a 3-year panel (2026-2028) where the load shape for each calendar month is a random selection from the corresponding donor months. The selected month from the bootstrap draw is consistent across balancing authorities to account for common shocks. We assign the load shape to our eGRID geography by using the shape of the most represented balancing authority within the eGRID region.[16] Lastly, each hourly load factor is then multiplied by the annual electricity demand number to yield the hourly load.

For each of our demand scenarios, we use reported power and operational dates for all operational and proposed data centers from CleanView, the same source we use for our ex-post analysis. We use 100% of reported data centers operating at 100% capacity as our highest case. For the two mid-range cases, we allocate compute across hours as before using estimated market shares and

[14]Note that this methodology does not account for data center demand crowding out other sources of electricity demand growth.

[15]For the eGRID regions that span multiple EMM regions, we use the sum of demand over the underlying regions for the growth rate.

[16]The majority of eGRID regions are largely or entirely comprised of one balancing authority.

utilization rates from Bloomberg/DC Byte. Unlike before, there is an additional source of uncertainty in data center demand. There is significant uncertainty in which data centers will actually be constructed. There have been a number of reports of data center developers submitting duplicate or speculative bids and many operators have constructed (or plan to) on-site primary and back-up generation. To represent this uncertainty, we first scale down the data center power capacity by 40% for all planned projects before constructing our mid-range scenarios. The low-end utilization case is a 50% reduction from the mid-range scenario. Hourly data center demand is added to the hourly demand without data centers constructed above when calculating the appropriate counterfactuals.

D.2 Forward Looking Supply Scenarios

Constructing forecasted thermal supply is straight forward. We use the EIA's Form 860m to gather the most up-to-date information on planned retirements, uprates, and derates to apply to our existing thermal panel at the monthly-level. We use the last available heat rate, capacity, and permit prices for each existing generating unit.[17] Form 860m also contains information on new planned thermal generation by month. We apply average heat rates from existing plants of the same generator type and of newer construction vintage (since 2017) to capture technological innovation. We apply the permit prices from units of the same subregion to planned units. Plants are derated and have stochastic outages as in our baseline approach.

In order to construct unit-level marginal costs, we require data on fuel costs. We retrieve the Henry Hub and WTI prices corresponding to the months drawn in our bootstrap, and for each day, we compute the deviation in observed price from the donor month's average annual natural gas or oil price. We multiply these coefficients by the annual natural gas and petroleum prices for electric power projected in the 2025 AEO. We apply the state-specific markups to these daily natural gas and oil prices to capture the time-invariant spacial differences in prices faced by our panel of thermal units. Coal and coke plants receive the same treatment, utilizing the fuel cost data

[17] We ignore unit degradation over time. Given our forward looking scenarios is over only a three-year period, we don't expect unit heat rates to change much over this time. To the extent that they do, our results will tend to under-estimate costs and emissions across thermals.

from the EIA Form 923 that corresponds to each donor month and projected coal prices from the 2025 AEO.[18]

For renewables, we use data from the EIA-930 to estimate regional month-by-hour capacity factors for installed solar and wind. We apply these hourly capacity factors for the forward looking years using the same bootstrap draw used for estimating the demand profile. We then apply these capacity factors to the total amount of installed wind and solar from existing units and planned units in EIA-860m.

We perform a similar exercise for hydroelectric units. We create average hourly generation profiles for each month by using hourly data from the EIA-930. For each subregion-hour within a donor month, we compute the average share of that month's total hydro generation that is utilized. We then apply those shares to the total monthly output of each hydroelectric facility, as reported in the EIA-930, for that donor month.[19] We use average hourly capacity factors by subregion for planned hydro units. We take this approach in order to preserve some of the complex seasonality and environmental constraints inherent to hydro operations without explicitly modeling reservoir shadow prices as in Bushnell (2003).

For nuclear, cogeneration plants, and other, less common generation technologies, such as geothermal plants, receive the same treatment we use the EIA-923 to compute average monthly capacity factors. We then assume all of these units generate at that fraction of their nameplate capacity for the given month.[20] While output from nuclear plants does show some variation across months due to factors such as maintenance outages/derates, their hourly production tends not to fluctuate, as it can be quite costly for these units to ramp up and down. Cogeneration plants are typically used in industrial processes, and thus typically do not respond to wholesale market signals. Other generation types, such as geothermal, tend to operate similarly to nuclear plants,

[18]The AEO does not include projected coke prices for delivered power. Because coal and coke are substitutes, the price of coke is highly correlated with coke. We observe that over our sample, the cost for coke is, on average, $0.40 per MMBTU higher than coal. Thus, we use the price of coal with a $0.40 markup for the small number of coke units in our panel.

[19]If the shares work out such that a facility ever produces above its nameplate capacity, we instead force that facility to have uniform output for all hours within a month.

[20]For cogeneration units, we first compute a monthly capacity factor, and multiply this by the average fraction of their reported net generation that is sold to the grid –a number that is reported annually in the EIA-923 data.

or are so uncommon (e.g. wood-burning generators), that the benefit to directly modeling their supply decision in a more detailed manner would provide little-to-no benefit to our analysis.

Along with solar, grid-scale storage capacity is expanding at a rapid pace. However, storage is unique in its role as an electricity arbitrageur. Batteries are constrained by their state-of-charge, round-trip efficiency, and rated power capacity (their upper limit on charging and discharging), rather than by fuel costs or daily weather patterns. Thus, we cannot treat it similarly to the technologies described above.[21] We treat the battery fleet in each subregion as a single price-taking agent who seeks to smooth the daily net load profile by charging in low net load periods and discharging during periods of high net load.

We construct the aggregate battery fleet in each subregion by combining data on the existing fleet with planned additions from EIA Form 860m. Planned additions report the rated power capacity but not the energy capacity, or duration of time the battery can continuously discharge at their maximum power capacity before having to recharge. To handle this, we probabilistically assign a duration of two or four hours.

We assume the objective of this subregional battery operator is to flatten the daily net load curve rather than to arbitrage prices across time. However, because we have no intra-regional transmission constraints, our wholesale prices will be convex with respect to net load. Thus, this objective function mirrors that of a profit maximizing agent. In addition, we assume this battery operator has perfect foresight over the net load within a 24-hour period.[22] The battery operator's "peak-shaving problem" is described as follows:

For each day, d, the operator observes hourly net load, d_t and chooses x_t and y_t, charge and discharge, to minimize the peak net load and maximize the net load trough. Formally, for each

[21] Hydroelectric resources also act as arbitrageurs, shifting energy delivery to periods of highest value. However, unlike battery storage which typically operates on a short-term, typically daily, cycle, Hydro scheduling is governed by additional intertemporal constraints spanning seasons (e.g., reservoir management and environmental flows). Modeling these constraints are beyond the scope of this paper. See Bushnell (2003) for further discussion.

[22] At the time of solving for battery charge and discharge, net load is equal to our forecasts for regional demand minus solar, wind, hydro, nuclear, cogeneration, and other thermal plants.

hour, t:

$$x_t = \min(K, \max(0, \underline{d} - d_t)), \tag{6}$$

$$y_t = \min(K, \max(0, d_t - \overline{d})), \tag{7}$$

where we use binary search to numerically solve our peak and trough thresholds, $\underline{d}$ and $\overline{d}$, subject to a daily energy balance constraint:

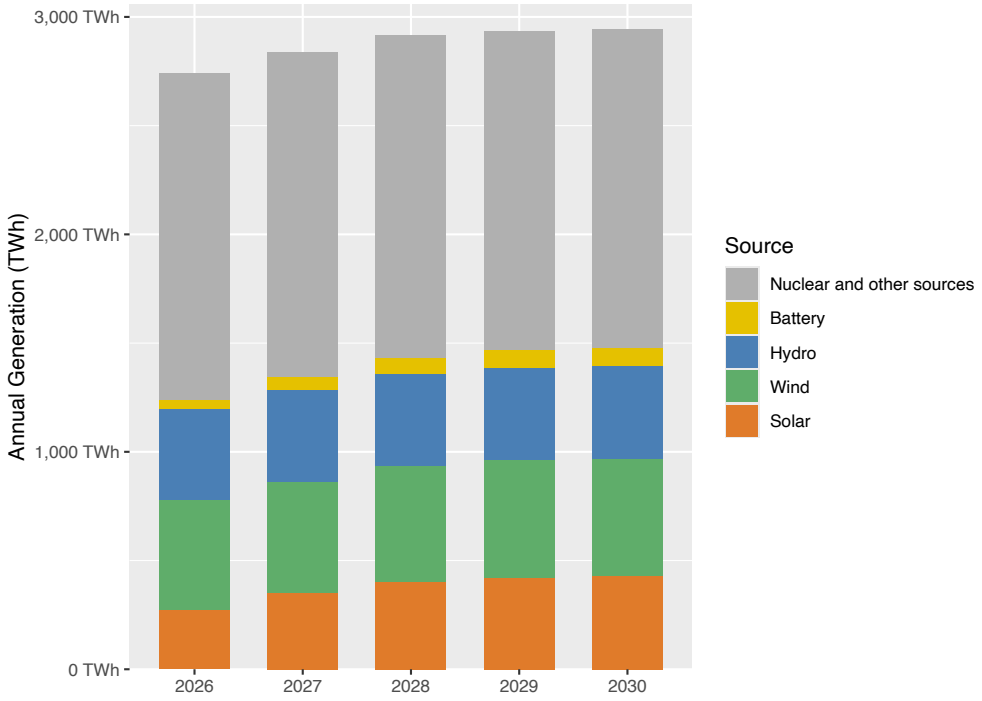
$$\begin{aligned} \sum_{t=0}^{24} y_t &= \eta \sum_{t=0}^{24} x_t, \quad \text{where} \\ \sum_{t=0}^{24} y_t &\leq E \\ \eta \sum_{t=0}^{24} x_t &\leq E \end{aligned} \tag{8}$$

Equation 8 ensures that the battery operator charge and discharge is exactly equal and does not exceed it total energy capacity (in MWh), E. The parameter η represents the round-trip efficiency of batteries, which we set at 0.85.[23]

Figure A10 shows the supply from non-thermal generators in each year used in our forecast. The majority of this generation comes from nuclear and other sources (predominately nuclear) though solar, wind, and hyrdo all contribute meaningfully. The majority of the growth comes from increased solar generation.

[23]Batteries lose some energy when charging and discharging. This value is typically 0.8 to 0.9 for grid-scale storage.

Appendix Figure A10: Non-thermal supply forecast



E Appendix: Cost Shift to Residential Price

In this Appendix, we consider how the additional electricity demanded from data centers will impact consumer bills for a representative utility that recovers both variable costs it incurs and total system fixed costs through a volumetric rate. Let TFC be the total fixed costs and $AVC(D)$ be the average variable costs (per kWh of delivered electricity) associated with delivering D kWh of electricity.[24] If total electricity demanded without data centers is D and the utility meets its revenue requirement with a linear price schedule, the price of electricity delivered to end use customers is given as

$$p^{-DC} = \frac{TFC}{D} + AVC(D).$$

Now consider the effect on prices for other consumers when data center demand increases total electricity demanded by ΔD. If data center demand does not increase the fixed costs of the

[24] Technically, it maybe more appropriate to consider TFC and VC as the revenue requirement from the load serving entity coming from fixed costs rather than simply the fixed costs themselves.

system at all, then the price of electricity will be given as

$$p = \frac{TFC}{D+\Delta D} + AVC(D+\Delta D).$$

This additional demand will have two offsetting effects. First, it will lower the average fixed costs, putting downward pressure on electricity rates by spreading fixed costs over more demand. Second, it will increase average variable costs as average variable costs increase when demand increases ($AVC'(\cdot) > 0$). This additional demand will lower bills for other end-use customers if $p \leq p^{-DC}$ or

$$\frac{TFC}{D+\Delta D} + AVC(D+\Delta D) \leq \frac{TFC}{D} + AVC(D)$$

Define the percent increase in electricity demand as $d = \frac{\Delta D}{D}$ and the percent increase in average variable costs as $v = \frac{\Delta AVC(D)}{AVC(D)}$ then prices will be lower after datacenter demand if

$$\frac{TFC}{D}\left(\frac{1}{1+d} - 1\right) \leq AVC(D) \cdot (1 - (1+v)).$$

Define the share of expenses from fixed costs as $s = TFC/(TFC + d \cdot AVC(D)$. Rearranging and substituting in s gives

$$\left(\frac{s}{1-s}\right) \leq \frac{v}{d}(1+d).$$

Note that $\eta \equiv d/v$ is the elasticity of electricity supply to the average variable cost. Solving for s gives

$$s \geq \frac{(1+d)}{\eta + (1+d)}.$$

The above equation gives a threshold criteria for the minimum share or residential rates coming from fixed costs that are needed in order for additional demand to lower rates.

F Appendix: The Social Planner's Problem

For each hour t, we solve a national dispatch with no regional constraints:

$$\begin{aligned} \min_{q_{it}} \quad & \sum_{i \in I} mc_{it} q_{it} \\ \text{s.t.} \quad & \sum_{i \in I} q_{it} = \text{DC Load}_t^{\text{national}} \\ & 0 \leq q_{it} \leq K_{it}^{\text{idle}} \quad \forall i \in I. \end{aligned}$$

That is, we minimize the total cost of generation required to meet total data center demand, $\text{DC Load}_t^{\text{national}}$, subject to a generator capacity constraint, K_{it}^{idle} —the capacity of generator i not dispatched in the no-DC counterfactual. For the ex-post period, this means we are solving over one of the counterfactuals which originally subtracted assumed data center demand from the observed net generation of our thermal panel. In the forecast period, since we build net load from the ground up, this means solving the planner's problem from the dispatch solution that does not include data center demand.