



Federal Reserve
Bank of Dallas

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

Owen Kay, Robert Reaser and Reid Taylor

Working Paper 2606

March 2026 (Revised October 2026)

Research Department

<https://doi.org/10.24149/wp2606r1>

Working papers from the Federal Reserve Bank of Dallas are preliminary drafts circulated for professional comment. The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Dallas or the Federal Reserve System. Any errors or omissions are the responsibility of the authors.

Processing Power: The Effect of Data Centers on Wholesale Electricity Markets*

Owen Kay[†] Robert Reaser[‡] and Reid Taylor[§]

March 17, 2026

Revised: October 9, 2026

Abstract

Data centers are reversing two decades of flat U.S. electricity demand. We estimate their effect on wholesale electricity markets using an hourly, unit-level dispatch model in which 23 U.S. regional markets trade over capacity-constrained interfaces. Data centers raised wholesale prices by 6.9 percent nationally by 2025 and by 15.0 percent in northern Virginia; under our preferred scenario, the national effect reaches 12.3 percent by 2028. Trade spreads these effects to regions hosting few data centers. Roughly half the increase in generator payments is a transfer to inframarginal generators. We derive sufficient statistics for when load growth lowers retail rates.

JEL codes: L94, P18, Q41, Q42, Q48

Keywords: electricity prices, energy, data centers, artificial intelligence

*[Ⓡ] 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 the NBER EEE Spring Program, Electricity Camp in the Rockies, CAISO, the Dallas Fed, UC Davis, and UC Santa Barbara. We are especially grateful to Mar Reguant for a very useful discussion of this paper. 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.

[†]Owen Kay, Williams College, email: owen.a.kay@gmail.com.

[‡]Robert Reaser, University of California Davis, email: rreaser@ucdavis.edu.

[§]Reid Taylor, Federal Reserve Bank of Dallas, email: reidbtaylor@gmail.com.

1 Introduction

After nearly two decades of flat electricity consumption in the United States, growth in electricity demand has re-emerged as a defining feature of power markets, in large part due to the rapid expansion of data centers for 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 can be constructed quickly relative to the pace of generation and transmission investment. Concerns over affordability have contributed to opposition to data centers across the political spectrum with states as politically diverse as Texas and New York calling for pauses or moratoriums on new data center construction. However, a growing body of work has argued that demand growth can, in fact, lower electricity bills for residential customers.

The economics of electricity markets suggests that the effect of a large increase in demand for electricity on prices is ambiguous. Consumer retail prices reflect both the short-run marginal cost of generating an additional kWh of electricity, as well as the costs of transmission lines, substations, distribution network infrastructure, and a variety of other costs traditionally considered to be fixed, at least in the short run. If additional demand can be met with existing grid infrastructure and no additional fixed costs, total fixed costs are spread over a greater number of kWh, decreasing the average fixed cost and putting downward pressure on customer rates.¹ On the other hand, an upward sloping short-run marginal cost curve for electricity means that additional demand will be met by higher cost power plants on the margin, putting upward pressure on wholesale prices and electricity rates. The ultimate effect of electricity load growth on retail prices thus depends on the relative magnitude of these effects and their respective weight in the end-users bill.

In this paper, we quantify the short-run effect of existing and proposed data center electricity demand on electricity prices in the US. To do so, we first use a detailed structural dispatch model

¹The opposite phenomenon is known as the “utility death spiral” by which decreasing load, due to increasing residential rooftop solar as an example, leads to a higher average fixed cost for remaining users, leading more users to shift off of the grid and continue the cycle (Davis and Hausman 2022).

to estimate the increase in wholesale prices and variable electricity costs caused by the increase in demand related to data centers. Using generator-level characteristics for the universe of grid-level power plants, we construct market-specific supply curves for each hour in each geographic subregion of the contiguous 48 US states over the recent past (2021-2025) and near future (2026-2028). We then solve for equilibrium prices and dispatched quantities with and without electricity demand from data centers accounting for plant outages, generator technical ramping constraints, and inter-regional trade flows. Because data center utilization is not directly observed, we evaluate a range of potential data center capacity and utilization scenarios calibrated to market-level aggregates by data center type. Comparing these counterfactuals allows us to estimate how much data centers have already increased wholesale electricity costs and how much these costs might increase under variable build-out scenarios in the future in different locations across the US.

Our structural approach is motivated by two characteristics of electricity markets which make estimating the effects of load growth on prices challenging. Electricity grids are highly interconnected nodal networks, meaning demand shocks in one location are likely to have spillover effects as generators both within region and in nearby connected regions can respond.² Accounting for trade flows when modeling load growth is especially important as failing to do so may overstate the availability of idle generation that would serve load in connected regions, as well as overstate the effect of data center arrivals by not dispatching cheaper idle foreign generation. Relatedly, the interconnected nature of the electricity grid also makes it difficult to find suitable controls for the use of reduced-form causal inference methods. Furthermore, because short-run supply curves are highly non-linear and convex, even well identified estimates based on historical data may not support credible extrapolation outside the study period and geography. Our model reproduces observed market outcomes closely, matching both unit-level generation and market-level prices with high correlations and observed regional trade flows.

Several key findings emerge. First, the realized effect is already economically meaningful and highly concentrated. By 2025, data centers raised the load-weighted national wholesale price

²Trade flows are large in magnitude. In 2025, 12% of thermal generation was traded across grid regions.

by 6.9 percent under our preferred scenario, increasing to 12.3 percent by 2028. The nationwide average effect masks considerable geographic heterogeneity driven both by the geographic concentration of data centers and the varying elasticity of electricity supply in different regions, either from local production or trade. As of 2025, under our preferred utilization case, Northern Virginia, home to the densest concentration of data centers in the country, experienced an increase of \$5.22 per MWh (15.0 percent), and the Pacific Northwest experienced an increase of \$5.68 per MWh (15.4 percent). At the other end, effects in Florida and parts of the interior South are as low as \$0.42 (1.0 percent). Due to trade flows, regions such as California with low data center demand growth are still estimated to have sizable effects from data centers driven through foregone imports as cheaper generation is used to meet data center demand locally in connected regions or exported from California during solar-rich hours rather than consumed locally, spreading the incidence of the data center build-out well beyond the communities and regions hosting the facilities.

Furthermore, a significant portion of the increase in payments to generators as a result of the higher wholesale prices is a transfer from existing customers. We use the unit-level hourly dispatch instructions and market prices from our model to decompose who bears the increased costs. Over 2021–2028, our preferred build-out scenario raises total payments to generators, valued at hourly market prices, by roughly \$186.7 billion.³ Of this amount, \$94.5 billion, or roughly 50%, accrues as inframarginal rents on generation that would have been produced anyway. The remaining 50% is due to the cost of supplying the additional electricity brought online to meet the data center demand plus negligible rents for the incremental generators.

We present a number of alternative analyses and robustness checks for the model. Notably, in addition to calculating pricing effects from data center demand, we calculate additional equilibrium outcomes for greenhouse gas emissions showing that in addition to increasing wholesale prices, data center electricity demand significantly increases greenhouse gas emissions. We also test the sensitivity of our wholesale market effect to alternative renewable build-out and coal retire-

³Because this figure values all generation at the contemporaneous market price, it measures the change in the spot-valued market-wide energy bill rather than realized cash flows. Long-term contracts, financial hedges, and cost-of-service recovery attenuate and delay the pass-through to both generator revenues and retail rates which we discuss in Section 5.

ment scenarios and endogenous entry of new power plant. Furthermore, we compare the observed location of data center compute against the geographic allocation a planner minimizing short-run operating cost would choose, allowing data center compute to be freely reallocated across regions hourly. This geographic flexibility of electricity demand acts as a form of “virtual transmission” providing meaningful cost reductions.

Finally, we use the predicted change in variable costs from our model to evaluate the conditions under which load growth could decrease retail prices for customers. Using a framework of a utility using a volumetric rate to recover both variable (energy) costs and fixed costs, we show that the ability of “fixed cost spreading” to offset energy cost increases depends on two sufficient statistics: the short-run electricity supply elasticity and the share of the per-unit retail rate dedicated to fixed cost recovery. Under the assumption that no additional fixed costs were incurred for the existing and planned data center build-out, we show that fixed cost spreading could have provided meaningful offsets to the upward pressure on retail rates in many regions, but likely fell far short in others.

We contribute to a growing literature on the economic and environmental consequences of electricity-intensive computing infrastructure. As projections for data center growth 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 expansion model to forecast long run effects 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 may impact which generation comes online and emissions profiles ([Knittel et al. 2025](#); [Ross and Ewing 2026](#)). In contrast, our analysis of siting decisions 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 Institutional Background

Electricity Markets

The US “grid” is comprised of three different grids, called interconnections: one in the eastern US, one in the western US, and one comprising most of Texas (ERCOT). These interconnections are largely independent from each other with very limited direct current (DC) transmission between them. Within an interconnection, electricity can be thought of as a commodity good subject to significant physical constraints. Grid operators must ensure the stability of the grid by matching supply and demand at every moment in time, accounting for the constraints of the transmission and distribution network and other grid infrastructure in order to avoid system-wide blackouts. New electricity load can be met by power plants elsewhere in the interconnection if adequate transmission capacity exists and the network is not congested. However, transmission is often at capacity, effectively fragmenting different geographies into separate, isolated sub-regions. The cost of serving new load is therefore geographically specific and dependent on whether demand occurs during periods when transmission infrastructure is congested or has excess capacity.

The market structures for electricity vary between regions. Historically, electricity production, transmission, distribution, and retailing were vertically integrated as the high fixed costs inherent in building out the grid led the industry to exhibit natural monopoly characteristics. A single utility owned power plants, transmission and distribution lines, and was responsible for ensuring that supply was equal to demand at each point in time, dictating production decisions for their portfolio of power plants. Starting in the 1990s, roughly two thirds of the contiguous US began to deregulate the generation sector, eventually organizing into one of seven wholesale electricity markets, where a separate Independent System Operator (ISO) or Regional Transmission Organization (RTO) manages a wholesale market to procure energy, capacity, and ancillary services.^{4,5}

⁴Many ISO/RTOs procure forward capacity through capacity auctions. Expected load growth can increase the prices in these capacity markets as expected demand increases. This mechanism is outside the scope of our paper which focuses on the energy portion of wholesale markets.

⁵Some vertically integrated utilities operate within ISO/RTOs and participate in the wholesale market, such as

Grid operators centrally manages which power plants produce in real-time by procuring electricity through auctions and dispatching the lowest bidders that are feasible for the grid with the price for energy being set uniformly at the price of the marginal producer.⁶ In regions that are part of wholesale markets, load serving entities (LSE) generally purchase electricity at wholesale prices and deliver electricity to consumers at marked-up retail rates.⁷

In contrast, in the southeast and parts of the west, power plant dispatch is still dictated by a vertically-integrated regulated utility. In these regions, the same utility produces electricity and delivers it to consumers, so market-wide wholesale prices do not exist. However, these utilities are regulated to supply electricity in the least-cost way such that the shadow price of the utility's internal least-cost dispatch algorithm (referred to as system lambda) is an economically meaningful value closely related to the wholesale price, measuring the marginal cost of producing an additional MWh of electricity.

The physical requirements that supply equals demand at each moment in time and the physical constraints of grid infrastructure influence the shape of short run electricity supply curves. The vast majority of electricity is produced from either natural gas, coal, nuclear, hydropower, wind, or solar. Individual power plants are limited by a maximum capacity they can produce at any time and marginal costs of production can vary dramatically between technology types. These characteristics lead to short-run supply curves that are typically upward sloping, convex, kinked at boundaries between technologies, and nearly vertical as the system approaches full capacity. This convexity implies that there is not a single supply elasticity, and price and cost increases from additional load can vary greatly depending on system constraints.

Retail rates that residential, commercial, and industrial consumers of electricity ultimately pay

Dominion Energy within PJM.

⁶ISO/RTOs dispatch according to Security-Constrained Economic Dispatch (SCED) algorithms that account for constraints on grid infrastructure, so higher marginal cost power plants are occasionally dispatched “out of merit order” ahead of cheaper alternatives to respect limits on transmission lines, substations, and other infrastructure. As a result, producers are paid a locational marginal price (LMP) set at each network node, conventionally decomposed into a system-wide energy component, a congestion component, and a loss component. The energy component is common across the market, so realized prices differ across nodes only through congestion and losses.

⁷Note that LSEs, utilities, and generators can engage in power purchase agreements that operate as forward contracts on electricity prices. However, any settlement of these payments happens between parties and outside of the wholesale market operation.

are set separately and heavily influenced by state-level policy and regulations. Rates are regulated by state public utility commissions, allowing for utilities to charge high enough rates to recover prudently incurred operating costs plus an approved rate of return on capital expenditures. Retail rates can be broken into two parts, the variable cost portion, which is largely comprised of the cost of generation, and the fixed cost portion which recovers the expenses of transmission, distribution, taxes and fees. As outlined by [Borenstein and Bushnell \(2015\)](#), many states undertook various degrees of “restructuring” of their electricity industries in the 1990s to move away from a fully vertically integrated model to introduce greater competition in electricity generation. However, the transmission and distribution systems remain monopoly controlled even in restructured states, and retail rates are still subject to state regulation and approval.⁸ For the purpose of our analysis, the main difference between restructured and non-restructured states relates to what comprises operating costs from electricity generation ([Borenstein and Bushnell 2015](#)). In restructured states participating in wholesale markets, the marginal cost of electricity, i.e. the wholesale price, determines the energy portion of operating costs passed through to consumers. In non-restructured states, the energy portion of consumers’ bills reflects the average variable cost of the utility’s dispatched fleet.

Data Center Electricity Demand

After decades of relatively flat demand growth, US electricity demand has begun to grow again, driven in large part by increased demand from data centers. While data centers have been present for decades, their increased prevalence and scale increased drastically with the introduction of AI models. Estimates of data center usage show that their electricity demand increased from 1.9% in 2018 to 4.4% of US total electricity demand by 2023. This number could increase further to 12% by 2028 ([Shehabi et al. 2024](#)).

⁸In states with retail choice, public utility commissions regulate transmission and distribution rates and oversee the procurement of default service, but do not approve the energy prices charged by competitive retailers. Most such states retain a regulated default supplier for customers who do not switch, however, in most of ERCOT retail choice is mandatory and no default supplier exists.

Data centers are siting across all regions of the continental US.⁹ To date, Virginia has led the way with the most operational data center capacity in the country. However, planned additions elsewhere, particularly in Texas, will substantially narrow Virginia’s lead by 2028.

Data centers consider a combination of factors when deciding where to locate. Historically data centers often located along major fiber optic cable corridors and near major urban areas to reduce latency and benefit from skilled labor markets. Electricity access is also a key consideration and has emerged as a primary limiting factor. Grid interconnection queues can often require projects to wait years 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 centers site in Appendix Table A1. This pattern is evident when comparing differences between counties with and without data centers. Counties with data centers have marginally higher electricity rates than areas without.

3 Data

Our analysis combines a number of data sources to construct a detailed model of electricity supply and data center electricity demand. On the supply side, we primarily use unit-level data on the thermal generating fleet, hourly data on regional demand and non-thermal generation and estimates of electricity transfer capacity between regions. For data center electricity demand, we use facility level data on existing and planned data centers.

3.1 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 pro-

⁹Appendix Figure A3 shows the geographic distribution and concentration of operational data center capacity in 2025, while Appendix Figure A4 shows operational and proposed data center capacity as of 2028.

duction and fuel use at nearly every conventional generator in the U.S.¹⁰ CEMS reports generator *gross* electricity production, which we convert into net production using plant-level gross-to-net ratios calibrated from EIA-923 data. 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. Power plants are frequently offline for either routine maintenance or unexpected outages. We follow Hausman (2025) and Ham et al. (2025) and uniformly derate generator maximum capacity by the planned outage rate and stochastically apply unplanned outages at the unit-day level, both calibrated to observed rates.¹¹ We use the EPA’s eGRID subregion designation to separate generators into regional markets.¹²

We build upon 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

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

We calculate the heat rate for generator i , $HeatRate_i$, from the observed ratio of total annual fuel consumption to annual net generation.¹³ Fuel cost estimates vary by type. For coal- and coke-fired generators, the variable fuel cost component is constructed from EIA data on state-specific monthly fuel costs paid by power plants. For natural gas and oil-fired generators, we apply a facility-by-year (or state-by-year, when facility-level data is not reported) specific markup

¹⁰CEMS 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 power into electricity markets.

¹¹Unlike Ham et al. (2025) which used a common derate factor across all regions, we use CAISO’s Resource Adequacy maintenance and forced-outage rates by unit-type in CAMX, NWPP, AZNM, RMPA, and ERCT to better reflect the fleets operating in the Western U.S.

¹²We further split generators located in the Virginia/North Carolina (SRVC) subregion to better capture important dynamics surrounding the large concentration of data centers located in Northern Virginia. We detail this split below when we discuss our construction of transmission interfaces between subregions. For plants that are not present in the eGRID data, we locate generators into subregions according to their location and balancing authority.

¹³We censor anomalous heat rates below 4 MMBtu per MWh and above 50 MMBtu per MWh.

computed from EIA-923 data on average fuel prices paid by power plants to daily fuel spot prices (Henry Hub and West Texas Intermediate, respectively).¹⁴

We account for relevant environmental trading program permit prices using data from the EPA, the Regional Greenhouse Gas Initiative (RGGI), and California Carbon Allowances. Lastly, technology-specific variable O&M costs are calibrated from the EIA based on [Hausman \(2025\)](#). All dollar values are inflation adjusted to 2024 values using the total CPI index.

For our forward looking analysis, we treat our supply forecast as exogenous given the relatively short time frame of our sample window. Lengthy interconnection queues for new supply, often greater than 5 years, mean that the new supply added from 2026–2028 is likely not in response to the sudden, unexpected boom in demand from AI.¹⁵ We incorporate announced and expected power plant additions, retirements, derates/uprates, and fuel-cost paths from EIA-based inputs as well as proprietary data from YES Energy. For new plants, we assign the technology-specific 25th percentile heat rate from existing plants constructed after 2017 to account for technological advancement.¹⁶ We continue to apply stochastic unplanned outages and maximum output for new plants is set using 90% of the announced nameplate capacities. Existing plants that are not uprated or downrated are treated the same between ex post and forward looking periods.

3.2 Trade Interfaces and Import Capacities

We utilize two sources of data to construct trade interfaces between neighboring subregions and import capacities for each individual subregion. Our primary source is the North American Electric Reliability Corporation’s (NERC) 2024 Interregional Transfer Capability Study (ITCS). As part of this study, NERC assessed the total transfer capability between neighboring transmission planning regions and estimated their total import capacity for 2024 and 2025. These regions largely

¹⁴In several regions, most importantly New England and New York, delivered gas prices can deviate substantially from the Henry Hub benchmark, particularly in the winter months when pipeline capacity tends to be tighter. In these regions, we replace the markup with a seasonal ratio of delivered to Henry Hub prices (Appendix C.1).

¹⁵In later robustness checks, we test this assumption and show the sensitivity of our results to endogenous supply growth and power plant retirements.

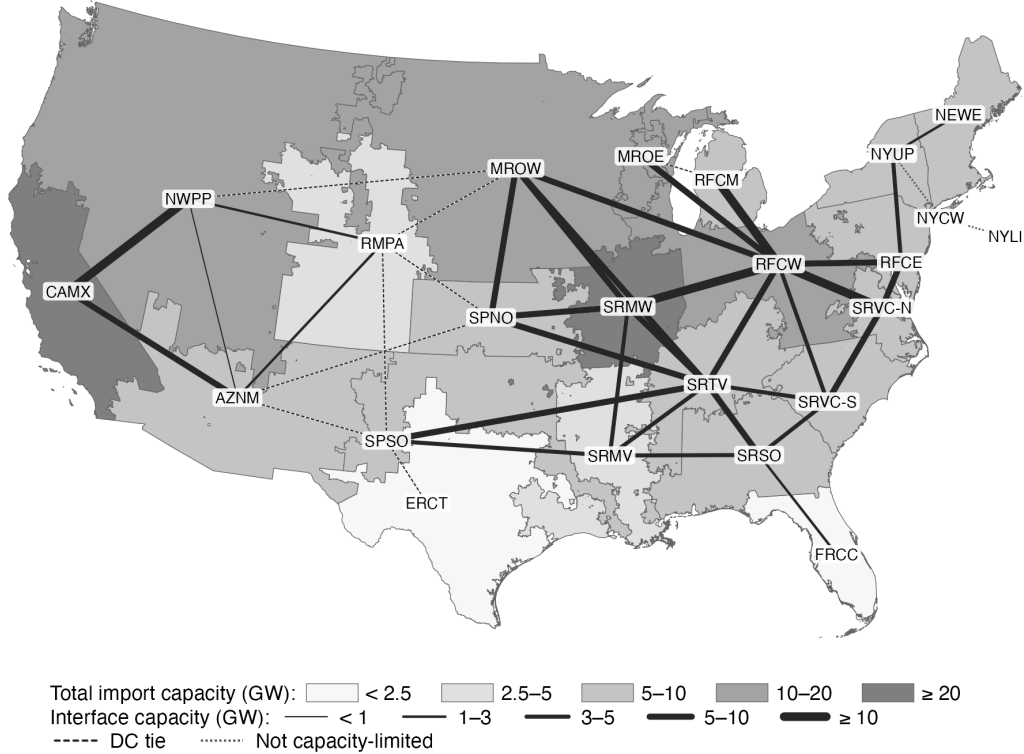
¹⁶A lower heat-rate means that the unit requires less fuel per MWh produced and is more fuel efficient.

map to our own eGRID subregions, allowing us to approximate their own constraints.¹⁷

Where eGRID and ITCS boundaries align, we use the ITCS Summer 2024 and Winter 2024/2025 capacities directly and assign shoulder seasons the simple average of the two. When subregions lack a direct mapping to an ITCS planning area, we supplement the ITCS with the 99th percentile of observed hourly interchange between the relevant balancing authorities from EIA Form 930. Finally, we cap each subregion’s total imports at its ITCS import capability. In total, we construct 42 interfaces between subregions. We display the resulting interfaces and import capacities in Figure 1, and Appendix C.4 details the mapping. Because moving power across interfaces is costly, we also price flows across interfaces with a hurdle rate, which we motivate and estimate in Section 4.2.

¹⁷Figure A5 displays the ITCS transmission planning areas, their transfer capabilities, and approximate import capacities.

Figure 1: Trade Interfaces and Total Import Capacities Between eGRID Subregions – Winter



Notes: This map displays the interface capacities between eGRID subregions as well as their total import capacities for winter, in GW. In summer and shoulder seasons, the total interface and import capacities tend to fall. Interface and import capacities are constructed from NERC’s 2024 ITCS report and supplemented with observed hourly interchange data from EIA Form 930. In total, there are 42 interfaces between subregions.

3.3 Residual Demand Data

3.3.1 Ex Post Periods

We construct each subregion’s residual demand, the load that is served by dispatchable thermal generation and inter-regional trade, by subtracting the observed supply of *inframarginal* generation from observed total demand. Our approach differs from [Hausman \(2025\)](#) and [Ham et al. \(2025\)](#), who use the observed thermal supply in a subregion as the implied residual demand. This method holds trade across subregions fixed at the observed level, even in the counterfactuals. Instead, we develop residual demand directly from demand data and model flexible interregional

trade using interfaces, which allows trade flows to vary across counterfactuals.¹⁸ To get hourly subregion electricity demand, we source data from the EIA’s Form 930, which reports demand at the sub-balancing authority level. Almost all sub-balancing authorities uniquely match to an eGRID subregion, allowing us to aggregate to an hourly subregion demand.¹⁹ From hourly demand, we subtract hourly inframarginal generation, which includes solar, wind, hydro, nuclear, geothermal, biomass, and cogeneration plants, using data from EIA-923 and EIA-930.²⁰

Three further adjustments ensure we model subregions on a common, physically consistent basis. First, we shift all subregion data to a common Eastern time so that each hour refers to the same physical hour everywhere. Second, using hourly interchange data from EIA-930, we subtract observed hourly net imports from Canada from adjacent subregions, treating cross-border flows as exogenous. This ensures that our residual reflects only generation that can be met within the continental U.S. Lastly, to handle the rapid expansion of grid-scale storage capacity in the latter part of our sample, particularly in the CAISO and ERCOT footprints, we model each subregion as having a single aggregate battery whose charge and discharge are chosen jointly with thermal output and interregional trade.²¹

3.3.2 Forecast Periods

Unlike the ex post analysis, where demand, including load from data centers, is observed, the forward-looking analysis requires constructing an hourly net load from total forecast demand and non-thermal supply forecasts. For each subregion-hour, forecast net load equals projected total demand minus projected output from the inframarginal resources. We construct hourly demand forecasts for 2026–2028 for the non-data-center portion of load by applying regional growth rates from the 2021 EIA *Annual Energy Outlook*, which predates the AI boom, to the ex post demand net

¹⁸Section 4.2 further discusses why this distinction matters when the counterfactual changes the level and location of demand.

¹⁹In some cases, sub-BA is not granular enough. We utilize data on retail sales from the EIA’s Form 861 to more accurately split RFCW and MROE, as well as to split balancing authorities NEVP and GRID between AZNM and NWPP.

²⁰Full details on the construction of hourly inframarginal generation are detailed in Appendix C.2.

²¹For details on the regional battery construction and how they enter the optimization, see Appendix D.3.

of data center load. A month-level block bootstrap of observed 2021–2025 load shapes generates realistic hourly load factors. We then add data center load, subject to our data center utilization assumptions, to model our counterfactual scenarios.²²

3.4 Data Centers

We obtain data on the characteristics of existing and planned data centers from S&P Global 451 Research’s Data Center KnowledgeBase (DCKB). These characteristics include the precise geographic location, data center type, net IT power capacity, the reported interconnection (utility) capacity, power usage effectiveness (PUE), a measure of data center efficiency, and operational dates. We compute the maximum potential load for each data center as net IT power (in MW) times PUE. PUE captures the additional power draw, per MW of IT compute, required for cooling and power distribution overhead, and typically ranges from 1.3 to 1.5.

For our model of electricity markets, we require a measure of hourly data center power demand. There are, however, several sources of uncertainty in the power demand for data centers. First, the DCKB does not report net IT capacity for planned data centers, which we impute from their reported utility capacity.²³ Second, actual utilization factors are unknown. Although many data centers publicly state their intent for 100 percent utilization, industry reports show real-world utilization is much lower and varies by data center type. Third, some data centers have installed on-site “behind-the-meter” main or backup power generation. Information on the size and utilization of these sources is often not publicly reported. In our utilization scenarios, we are agnostic about whether reduced utilization is due to lack of computing needs or a switch to on-site generation, as the two have the same impact on the electricity grid.

We reflect this uncertainty by constructing four data center demand scenarios, summarized below in Table 1, that vary both the power capacity (the maximum power draw at a point in time) and the utilization rate (the percentage of maximum power used in a given hour). In the High

²²Full details on demand and supply forecasts are provided in Appendix F

²³For each data center type, we compute the mean ratio $\rho = \text{Net IT Power} / (\text{Utility Power} / \text{PUE})$ from the set of 2024 and 2025 data center vintages. We model the grid load for planned facilities as $\rho \times \text{Utility Power}$.

scenario, we assume 100 percent of installed power capacity is used in all hours. For the reasons noted above, this scenario is highly unlikely, but it serves as an upper bound for the analysis. For the remaining scenarios, 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 allocate a percentage of power from hyperscale, cloud, and neo-cloud type data centers equal to the relative market share of training data centers to every hour, as these have a 100 percent utilization rate,²⁴ and apply type-specific utilization rates to the remaining data center capacity. In our Mid-Range-Constant scenario, we assign this power equally across all hours of the day, while in our Mid-Range-Peak scenario, we vary daily electricity demand for each data center around its type's utilization rate in a load-following pattern constructed to mirror the patterns documented in [EPRI \(2026\)](#): consumption is highest from 11AM to 3PM, lowest from 10PM to 4AM, and equals the daily consumption of the Mid-Range-Constant scenario. Lastly, we evaluate a Low scenario which assumes 50 percent less power in each hour than the Mid-Range-Constant scenario.²⁵

²⁴While we do observe data center type in the DCKB data, we do not observe which data centers are dedicated to AI training versus inference, and these designations are not necessarily fixed. This motivates our decision to split the types most likely to host training (hyperscale, cloud, and neo-cloud) into training and non-training shares.

²⁵See Appendix [C.3](#) for more details on how we model data centers.

Table 1: Data Center Demand Scenarios

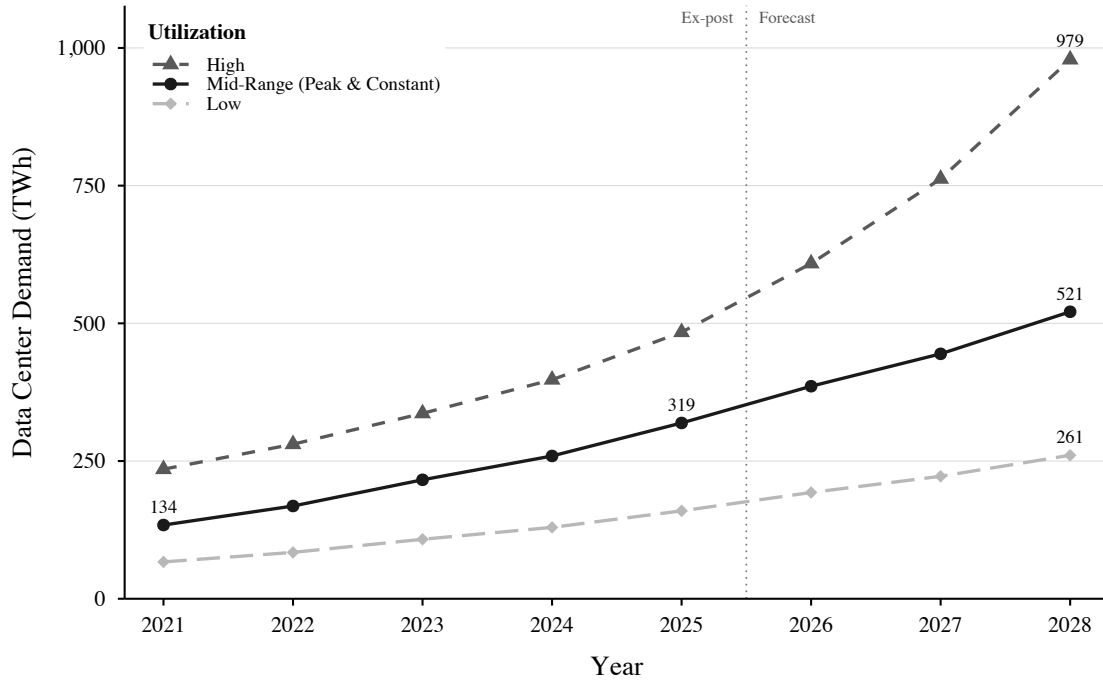
Scenario	Power capacity	Utilization	Intraday profile	Planned capacity (forecast)
High	100% of reported	100% in every hour	Flat	100% of announced
Mid-Range-Peak (preferred)	100% of reported	Training share at 100%; remainder at type-specific rates	Load-following	60% of announced
Mid-Range-Constant	100% of reported	Same as Mid-Range-Peak	Flat	60% of announced
Low	50% of Mid-Range-Constant	Same as Mid-Range-Peak	Flat	30% of announced

Notes: Power capacity is maximum potential load (net IT power $\times$ PUE) from S&P Global 451 Research’s Data Center KnowledgeBase. Utilization rates and training market shares by data center type are from Bloomberg/DC Byte; the training share applies to hyperscale, cloud, and neo-cloud facilities. Mid-Range-Peak and Mid-Range-Constant have identical daily energy and differ only in the hourly profile. The last column applies to facilities not yet operational at the start of the forecast period (2026–2028). New facilities ramp linearly to full capacity over 3 to 24 months, depending on type.

We evaluate the same four scenarios in both our ex-post and forward looking periods. However, to reflect the additional uncertainty from data centers not completing construction or connecting behind the meter, we scale down *new* data center capacity by 40 percent of announced capacity in both of the mid-range scenarios and Low scenario. Because construction timelines vary within and across data center types and campuses often scale multiple buildings over time, we assume that new facilities ramp linearly to full capacity over 3 to 24 months, depending on type.

Figure 2 reports the annual data center electricity consumption implied by each scenario. Under the mid-range scenarios, consumption rises from 134 TWh in 2021 to 319 TWh in 2025 and 521 TWh in 2028. We treat the Mid-Range-Peak scenario as our preferred central case, with the High and Low scenarios bounding the effects. All effects are reported relative to a no-data-center counterfactual in the same time period.

Figure 2: Annual Data Center Electricity Consumption by Scenario, 2021-2028



Notes: This figure shows the annual electricity consumption values, in TWh, that are assumed under our four data center scenarios from 2021–2028. Data center capacities are computed from S&P 451 Research’s Data Center Knowledge Database (DCKB). All values left of the dashed, vertical line are derived from operational data centers while those to the right come from projects that range from announcements to under-construction.

4 Structural Dispatch Model

4.1 The Dispatch Problem

We construct a model of short-run economic dispatch that replicates the grid operator’s cost minimization problem subject to a set of constraints. In each hour, we serve residual demand in each subregion with a least-cost combination of thermal generation, battery discharge, and imports from neighboring subregions. Generators are dispatched in order of marginal cost, conditional on their available capacity and physical limits on how quickly they can adjust their output between consecutive hours. Power flows between subregions up to the transfer capabilities of the interfaces that connect them allowing cheaper plants that would otherwise be idle to serve load in connected regions.

Formally, we solve:

$$\min_{q,f,x,y} \sum_t \left[\sum_{i \in I} \left(mc_{it} q_{it} + \rho_i (q_{it} - q_{i,t-1})^+ \right) + \sum_{e \in E} w_{e,s(t)} (f_{et}^+ + f_{et}^-) + \sum_r \kappa_r (x_{rt} + y_{rt}) \right] \quad (2)$$

$$\text{s.t. } \sum_{i \in I_r} q_{it} + y_{rt} - x_{rt} + \Phi_{rt} = d_{rt} \quad \forall r, t \quad (3)$$

$$0 \leq q_{it} \leq K_{it} \quad \forall i, t \quad (4)$$

$$q_{it} - q_{i,t-1} \leq R_i \quad (t \geq 2), \quad q_{i1} - q_i^0 \leq R_i \quad \forall i \quad (5)$$

$$0 \leq f_{et}^+ \leq \bar{f}_e^+, \quad 0 \leq f_{et}^- \leq \bar{f}_e^- \quad \forall e, t \quad (6)$$

$$\sum_{(e,\delta) \in \mathcal{J}_g} f_{et}^\delta \leq \bar{M}_g \quad \forall g, t \quad (7)$$

$$\sum_{e \in \mathcal{C}_k} f_{et}^{A \rightarrow B} \leq \bar{C}_k^{A \rightarrow B}, \quad \sum_{e \in \mathcal{C}_k} f_{et}^{B \rightarrow A} \leq \bar{C}_k^{B \rightarrow A} \quad \forall k, t. \quad (8)$$

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 hourly non-thermal generation. Coal and combined-cycle units pay a per-MW upward ramping cost, ρ_i , which we calibrate to the estimates from [Gowrisankaran et al. \(2026\)](#).²⁶ Generation from each unit is bounded above by available derated capacity, K_{it} . The constraint in Equation (5) limits a unit's change in output between consecutive hours to its physical ramp rate, R_i .²⁷ We solve the model at daily intervals jointly across all subregions, so that generator ramping links hours within a day and trade links subregions within each hour. Each unit's output in hour 0 of each day is anchored to its simulated output in the final hour of the previous day.²⁸ Each subregion's grid-scale storage fleet is represented by a single battery that charges x_{rt} and discharges y_{rt} , paying a throughput cost κ_r per MW. Discharging behaves like dispatchable generation, and charging adds to the subregion's load, so a battery can charge from imports and

²⁶[Gowrisankaran et al. \(2024\)](#) estimate a fleet-wide ramping cost of \$47.70/MW and \$38.60/MW for coal and combined-cycle units, respectively, in 2006 dollars.

²⁷We constrain up- and down-ramp by the unit's physical limits. Following the previous literature, we do not allow coal to ramp up or down more than 10% of its operational capacity between any two hours ([Gonzales et al. 2023](#)).

²⁸We include a six-hour lookahead period in each day's solve to ensure end-of-day dispatch is consistent with the following day's conditions and to avoid end-of-day behavior.

discharge to support exports.²⁹

Interfaces $e \in E$ govern trade flows between regions, with variables f_{et}^+ and f_{et}^- representing the forward and reverse flows running from origin region $o(e)$ to destination region $d(e)$. Net imports into each region are collected as

$$\Phi_{rt} = \sum_{d(e)=r} (f_{et}^+ - f_{et}^-) + \sum_{o(e)=r} (f_{et}^- - f_{et}^+).$$

Thus, each subregion's power balance (Equation 3) requires local generation and net battery discharge plus net imports Φ_{rt} to equal residual demand d_{rt} . We impose direction-specific limits on the amount of power that can flow across each interface, $\bar{f}_e^+$ and $\bar{f}_e^-$, and the combined flows across each shared corridor, k , are bounded by $\bar{C}_k^{A \rightarrow B}$ and $\bar{C}_k^{B \rightarrow A}$.³⁰ We also limit the total power subregion g can import to $\bar{M}_g$, where $\mathcal{J}_g$ is the set of interface-direction pairs (e, δ) whose flows enter g . Flows pay a hurdle rate $w_{e,s(t)}$ per MWh, which varies by interface and season $s(t)$.³¹

The model yields two main objects of interest: optimal dispatch quantities, q_{it}^* , for each generating unit, and the shadow value on the power-balance constraint, λ_{rt}^* , which is interpreted 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, λ_{rt}^* should be interpreted as a benchmark for the competitive wholesale price.

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 those 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 wholesale price.

Our counterfactual is best viewed as a short-run partial-equilibrium exercise that isolates how

²⁹The battery's power, energy, and state-of-charge constraints are given in Appendix D.3.

³⁰In some cases, overlap between subregions requires that we assign an interface e to the overlapping subregions. We cap their combined imports and exports at $\bar{f}_e^+$ and $\bar{f}_e^-$.

³¹Appendix C.4 describes the construction of the interface and import limits.

the marginal cost of serving load changes when data center demand is added or removed. Thermal dispatch, battery operation, and interregional trade are all re-optimized, while non-thermal generation, fuel prices, and generator availability are held at their baseline values. The effect of data centers on prices is:

$$\Delta\lambda_{rt}^\sigma = \lambda_{rt}^*(d) - \lambda_{rt}^*(d - DC^\sigma), \quad (9)$$

and we define effects on generation costs, payments, and emissions analogously. We aggregate these effects across subregions and hours using load weights.

4.2 Interregional Trade

The inclusion of transmission interfaces and import constraints yields three additional objects of interest. The multipliers on the directional interface limits in Equation (6), μ_{et}^+ and μ_{et}^- , equal the reduction in the total cost of dispatch from relaxing the limit on interface e by an additional MW of transfer capability in hour t . Similarly, ν_{gt} and ξ_{kt} are the shadow prices of relaxing a subregion's aggregate import capability (Equation 7) and a shared corridor's capability (Equation 8), respectively. The first-order conditions of the dispatch problem tie these shadow prices to the prices of the two regions an interface connects. For a flow from $o(e)$ to $d(e)$ on interface e in corridor k ,

$$\lambda_{d(e),t}^* - \lambda_{o(e),t}^* \leq w_{e,s(t)} + \mu_{et}^+ + \nu_{d(e),t} + \xi_{kt}, \quad \text{with equality whenever } f_{et}^+ > 0, \quad (10)$$

and symmetrically for flows in the reverse direction.³² When no limit binds, prices in connected subregions can differ by the hurdle rate, at most. A demand shock that raises the price in one subregion by more than the hurdle rate draws in imports from its neighbors, creating a spillover effect. This raises their prices until the gap returns to the hurdle rate or a transfer limit binds, at which point the shadow prices μ , ν , and ξ absorb the remaining difference.

³²Imports into California pay an additional per-MWh charge equal to an implied emissions rate times the California–Québec allowance price. We back out the emissions rate, by season and subregion (0.15–0.22 t CO₂/MWh), from the GHG component of CAISO real-time prices in hours with observed imports. We do not attribute imports to specific out-of-state generators, as CAISO does, because in our model renewable and hydro output enters as fixed net load rather than as dispatchable supply.

Prior work using similar dispatch models, such as [Hausman \(2025\)](#) and [Ham et al. \(2025\)](#), sets modeled regional residual demand for thermal generation equal to the observed thermal output. This mechanically holds net interchange between subregions fixed at observed levels. Given the object of interest is the gain in allocative efficiency from removing barriers to trade between regions and demand is held fixed, this modeling choice is appropriate. In our setting, however, the counterfactual involves a large demand shock that is concentrated in particular subregions. Holding trade fixed means that all data center demand must be met by generators within the same subregion. This would overstate the price effect in subregions, such as SRVC-N, that experience large demand shocks but which have large import capabilities with neighboring regions. It would also miss the effect of foregone imports in subregions, such as CAMX, where much of the price increase comes from foregone trade with neighboring regions using their cheaper generation to serve their own data center demand. On the other hand, aggregating subregions into larger markets with no limits on internal trade understates local price effects when interfaces would be binding between regions. Our approach falls between these two extremes.

Ramping rates and ramping costs matter for a related reason. Data center demand that follows a daily load shape, as in our Mid-Range-Peak scenario, must be met by units that can adjust their output within the day, which steepens the effective supply curve in shoulder hours relative to a static merit order.

4.3 Model Validation

The model recovers observed market outcomes accurately. Appendix Figure [A12](#) plots observed against 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, with off-diagonal points coming from our uniform capacity derate procedure. The correlation between simulated and observed unit-level generation at the hourly level is 0.634 (Appendix Table [A5](#)). Two features of the model cause unit-level generation to diverge. First, following [Hausman \(2025\)](#) and [Ham et al. \(2025\)](#), we uniformly derate

online generation to approximate planned outages and we draw unplanned outages stochastically, which will mechanically lower unit-level fit. In addition, adding ramping costs to a model where the counterfactual of interest removes demand requires that this number is expected to be less than one. Aggregating to the month and year level, the correlations rise to 0.784 and 0.867. This is further evidence that stochastic outages play a large role in the hourly-level correlations for quantities. Simulated prices track observed energy prices in the seven ISOs and RTOs in our sample, with correlations of 0.678 at the hourly level, 0.84 at the monthly level, and 0.929 at the annual level.³³ We provide further discussion of the generation and price fit in Appendix E.2.

Our no-data-center counterfactual changes demand heterogenously among the subregions we study, which means it is crucial to properly capture how each subregion responds to local demand shocks. We test this directly with the following regression:

$$G_{rt} = \beta R_{rt} + \alpha_{r,d(t)} + \gamma_{r,m(t),h(t)} + \delta_{c(r),t} + u_{rt}, \quad (11)$$

where G_{rt} is the hourly net generation of subregion r 's dispatched thermal units and R_{rt} is each subregion's hourly residual demand as defined above. We include subregion $\times$ date $\alpha_{r,d(t)}$, subregion $\times$ month $\times$ hour-of-day $\gamma_{r,m(t),h(t)}$, and interconnection $\times$ hour-of-sample (i.e. West, East, and ERCOT). Therefore, our coefficient β is identified only from variation in demand specific to each subregion—analogue to the shock a data center would impose on a subregion. It measures the share of this increase met by the subregion's own thermal generation, while $1 - \beta$ measures the share met by imports and storage. Without extending the model to include trade, β would equal one by construction. Our measure of residual demand is constructed by subtracting several separately measured sources of supply, and any errors in them would lead to attenuation bias (Hausman 2001). Following standard procedure in the electricity markets literature, we instrument residual demand with observed demand, as it is measured directly and considered perfectly inelastic in the short-run (Zivin et al. 2014). We also run a second regression that replaces G_{rt} with the thermal output of adjacent

³³Both series are winsorized at the 1st and 99th percentiles and Winter Storm Uri excluded (Appendix Tables A6 and A7)

subregions, while controlling for their demand.

Table 2 reports the results. We estimate that a subregion’s thermal units meet 64% of local shocks, the remaining 36% comes from outside of its own thermal fleet. Thermal units in adjacent subregions increase their output by 8% in response to the shock. Our model captures this pattern broadly, with 62%, 38%, and 10% for our respective estimates. Further, the observed data rejects $\beta = 1$ for all of the subregions in our sample. Notably, the own-thermal response is lowest in the subregions with the highest data center concentrations, such as SRVC-N, NWPP, and MROW. The model’s subregion-level estimates track the observed patterns in the data closely, with a high correlation across all estimated coefficients. As expected, our model tends to meet local demand shocks with more imports than observed, given the limitations to capturing all relevant frictions in the network.

Table 2: Who Serves an Increase in Local Demand? Data and Our Model

	Local demand shocks, 2021–2025	
	Data	Our model
<i>Panel A. All subregions: share of a local demand increase met by</i>		
the subregion’s own thermal units	0.64 (0.004)	0.62 (0.005)
other sources (imports, storage)	0.36	0.38
thermal units in adjacent subregions	0.08 (0.007)	0.10 (0.008)
<i>Panel B. Share met by the subregion’s own thermal units, largest host subregions^a</i>		
ERCT ^a	0.76	0.93
SRVC-N	0.31	0.24
NWPP	0.15	−0.14
MROW	0.48	0.35
RFCW	0.71	0.75
SRSO	0.84	0.63
SRVC-S	0.63	0.52
AZNM	0.33	0.32
SRTV	0.60	0.46
CAMX	0.43	0.37
Correlation with the data, other 13 subregions ^b		0.91

Notes: Instrumental-variables regressions of the subregion’s hourly thermal output on its residual demand. Each regions residual demand is instrumented with its observed demand. The estimation includes subregion×date, subregion×month×hour-of-day, and interconnection×hour-of-sample fixed effects. Standard errors are clustered by date. Row 3 regresses thermal output in adjacent subregions on the same demand shock, controlling for own-subregion demand.

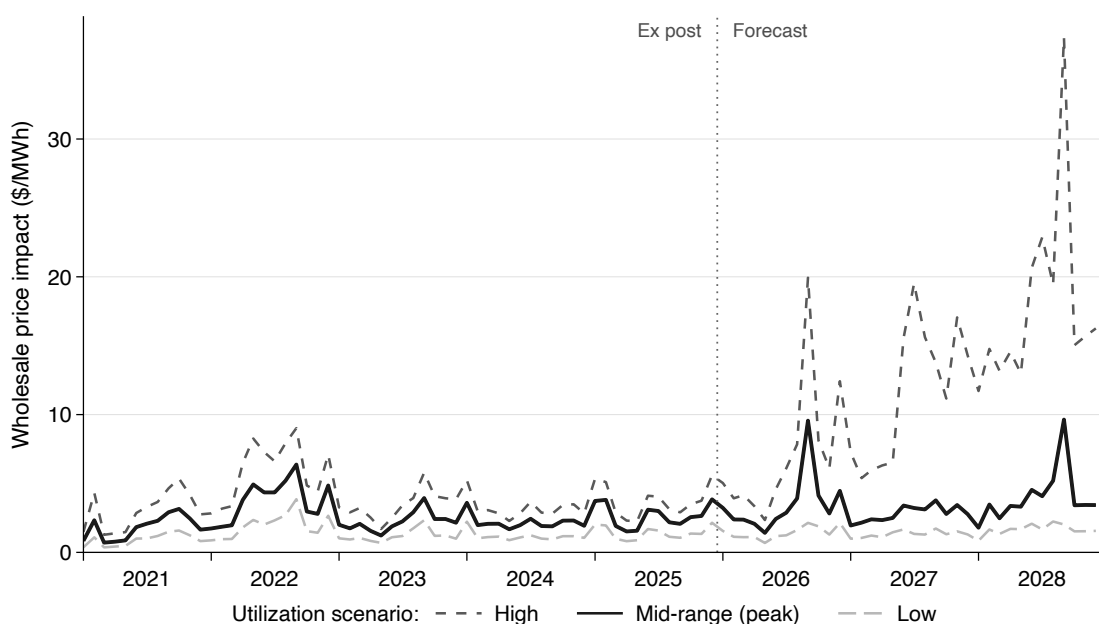
5 Wholesale Market Effects

5.1 Competitive Wholesale Price Effects

We begin our discussion of results by estimating the change in the wholesale electricity price due to data center build-out from 2021–2028 using our dispatch model. 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 for the price paid by utilities and load serving entities to procure electricity for retail customers.

We report our estimates of the nationwide load-weighted average monthly price effects of data centers under our four data center demand scenarios in Figure 3. Focusing first on 2023-2025, in our preferred mid-range scenario, wholesale prices were \$1.22 to \$3.94 more expensive, on average, than in the counterfactual world without data center load, varying across months. This translates to a 4.2% to 10.1% increase in the average wholesale price. The result is bounded above by the high-utilization scenario reaching \$5.79 (15.5%) and below by our low-utilization scenario with a minimum monthly average price increase of \$0.69 (2.3%).

Figure 3: Realized and Projected Average Monthly Wholesale Price Impacts, 2021–2028



Notes: This figure shows the load-weighted average monthly wholesale price increase by data center utilization scenario. Price increases are computed as the difference between each eGRID subregion's hourly wholesale price under observed and counterfactual simulations and then averaged to the monthly level. Both series are weighted by the no-data-center demand. Prices are in 2024 dollars per MWh.

As outlined in Section 4, the market supply curve is time varying, resulting in variation in the effect across time that is due to additional factors other than growth in data center demand that shift costs or steepen the supply curve. During the ex post period, the highest estimated price effects occurred during the summer of 2022 due to the large increase in natural gas prices caused by the Russian invasion of Ukraine. In the mid-range scenario, data center demand added \$6.37 to the nationwide average wholesale price at its peak in September 2022. As natural gas generators

are often the marginal generator and setting the market price, the large increase in input fuel costs resulted in a steeper supply curve and increased the effects on the wholesale price for the same amount of load increase from data centers.

Our estimates also reveal a notable seasonal pattern in the effect. The highest price effects are concentrated in the summer and winter months when demand and prices are already the highest. As data center load expands over the sample, the market clears on even steeper portions of the supply curve, requiring more expensive generation units required to run to meet the additional load.

Turning towards the forward-looking periods, our results show that under our middle and low utilization cases, price differences from data centers increase almost doubles from 2025 to 2028. The national average price effect averages \$4.01 (12.3%) higher by 2028 in the mid-range scenario while the low scenario sees a 5.2% increase. There is a noticeable sharp increase in the upper bound effect under our high-end scenario where all planned data centers are built and operate with 100% utilization. This is driven by the substantial acceleration in the pace of the build-out and the growth in planned projects. High scenario prices are an average of \$17.88 (55.2%) higher than a no-data center counterfactual. The growing divergence between the effects in the high- and mid-case highlights the role that the convexity of the supply curve and binding transmission constraints play in leading to non-linear price effects with respect to demand growth.

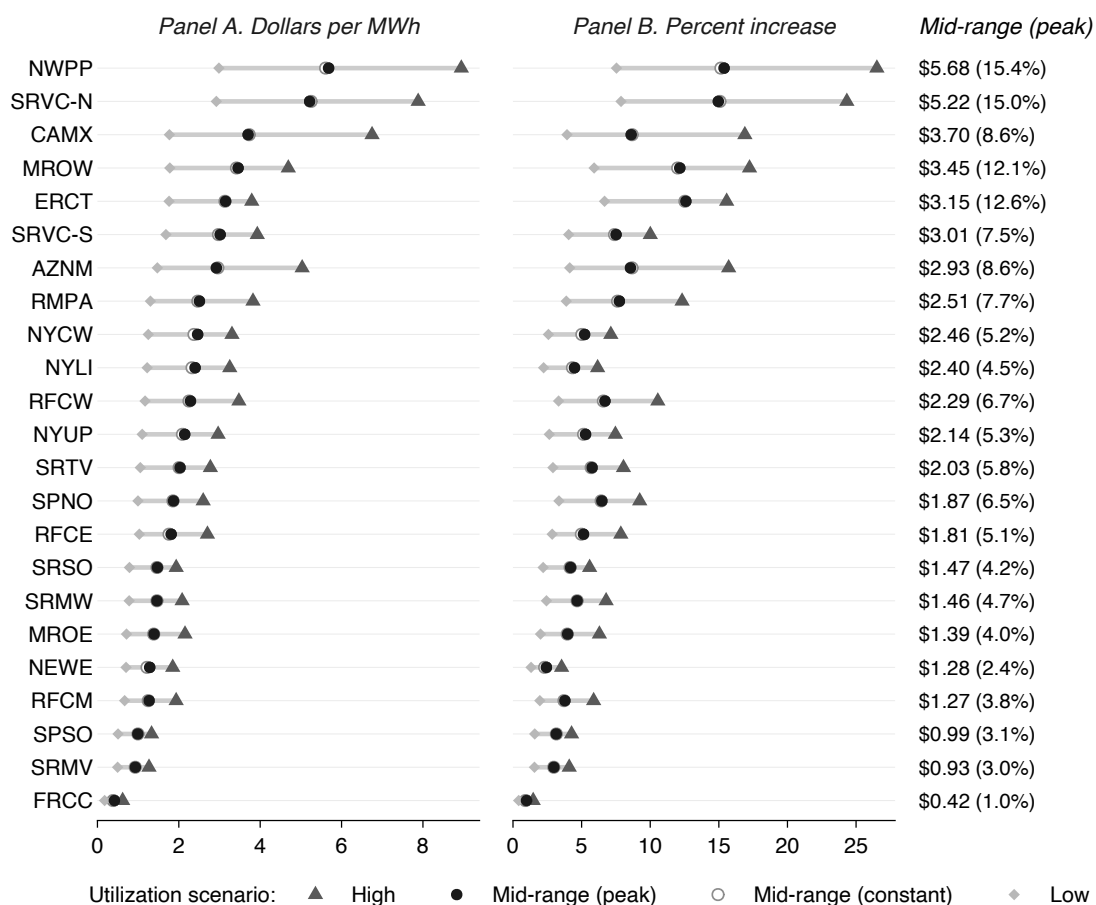
The distribution of data center load across the U.S. and the heterogeneity in regional power plant fuel source composition also means that there is substantial variation in results by region. In Figure 4, we show the average price impacts by eGRID subregion and utilization scenario for 2024 and 2025. The largest effects are concentrated in Virginia (SRVC-N) and the Northwest Power Pool (NWPP). Northern Virginia is home to “Data Center Alley,” the densest collection of data centers in the world. Northwest Power Pool (NWPP), which covers the majority of the Pacific Northwest, has the next largest effect from data centers despite not having as much data center driven demand growth as other regions. In our mid-range scenarios, we estimate data centers were responsible for a \$5.22 increase in competitive wholesale prices in SRVC-N, a 15.0% increase, and

a \$5.68 increase in NWPP (15.4%)

An important implication of the connected nature of the grid is that data center growth in one region can have spillover effects on to prices in adjacent regions through changes to trade. As an example, we estimate the third largest price effect (\$3.70 for 2024–2025 in the mid-range scenario) from data centers in the subregion covering most of California, CAMX, despite only modest growth in installed data center capacity. The price increases were instead primarily driven by foregone trade with the adjacent Northwest Power Pool (NWPP) and Arizona/New Mexico (AZNM) regions which, with their large growth in data centers, required local generation to meet this increased demand, making these cheaper resources not available for export.

ERCOT (ERCT), which is the grid covering the vast majority of Texas, by contrast has a smaller price effect despite containing the second-highest data center capacity behind Virginia (SRVC-N) and despite not being connected for trade to other regions. Data center load has added \$3.15 (12.6%) to wholesale prices, tempered by ERCOT's nation-leading supply build out, driven largely by renewable generation sources and battery storage.

Figure 4: Average Regional Competitive Wholesale Price Impact, 2024-2025

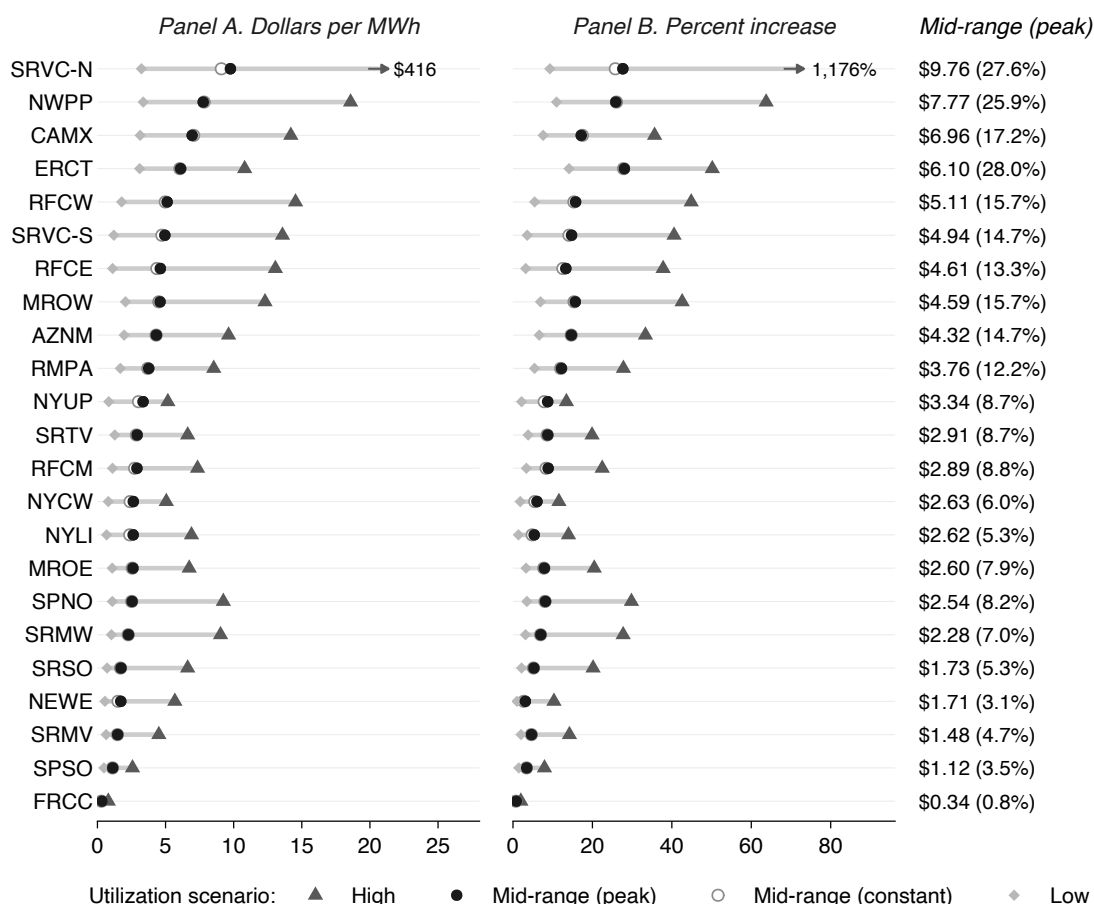


Notes: This figure shows the load-weighted average wholesale price impact of data centers by data center utilization scenario and eGRID subregion for 2024–2025. Price increases are computed as the difference between each eGRID subregion’s hourly wholesale price under the observed and counterfactual simulations and then averaged at the monthly level. Prices are in 2024 dollars per MWh.

As with our ex-post analysis, for our forward-looking results for years 2026–2028 we estimate substantial heterogeneity across regions driven by the concentrated nature of data center demand. Figure 5 presents the average regional wholesale price impacts that would materialize under our projected scenarios in 2028. Though the price impact of \$416.20 attributed to SRVC-N is an implausible magnitude in our high scenario, it highlights an important point. First, in this scenario, data center load growth out-paces local generation growth in SRVC-N. This excessive demand growth also causes trade interface constraints to bind, forcing SRVC-N to meet demand largely with local generation resources with the marginal generator being its most expensive generation

resource.³⁴ Thus, we interpret this scenario as a worst case scenario, and estimate the price effect if demand and supply were to materialize under this scenario while noting that market forces make this outcome highly unlikely.

Figure 5: Projected Average Regional Competitive Wholesale Price Impacts, 2028



Notes: This figure shows the load-weighted average wholesale price impact of data centers by data center utilization scenario and eGRID subregion for 2028. Price increases are computed as the difference between each eGRID subregion's hourly wholesale price under the observed and counterfactual simulations and then averaged at the monthly level. Prices are in 2024 dollars per MWh.

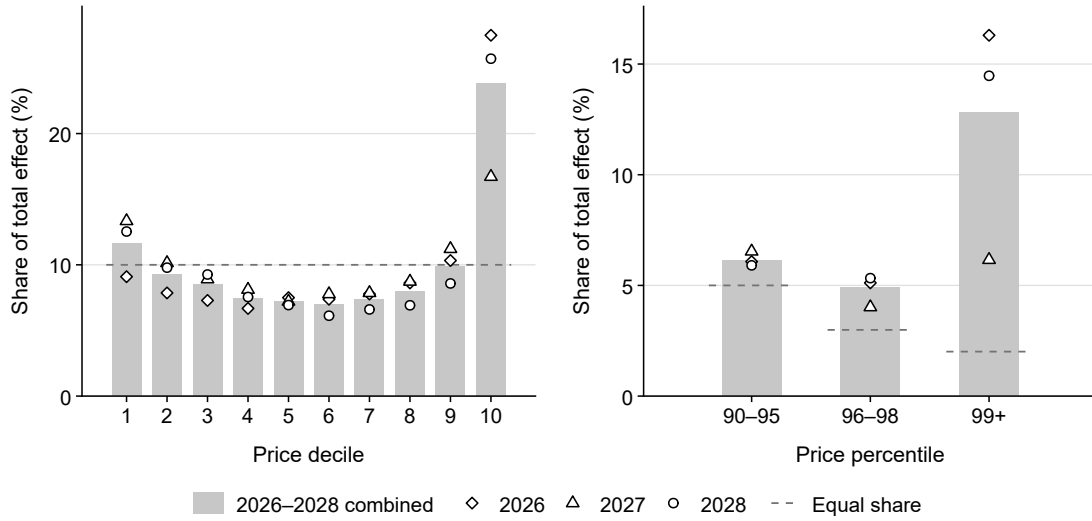
Our Mid-Range scenarios estimate more modest increases in data centers' effects in the highly impacted regions, such as SRVC-N and NWPP, relative to their ex-post impacts. SRVC-N sees price increases of less than \$4.54 while NWPP's price impacts increase by \$2.09. Importantly, the largest relative price changes from ex post to the forward looking periods occur in regions adjacent

³⁴In a substantial amount of hours for SRVC-N (96.7%) there is not enough supply available to meet demand in the forward looking periods under the high scenario, and no market-clearing solution is available. In this case, the market price is set by the marginal cost of the final dispatched generator.

to SRVC-N and NWPP due to the effect data center demand growth has on interchange between them. For example, the price impacts in CAMX increase by more than 88% from the ex-post to forecast period. CAMX does not see substantial data center growth in the forecast period, but sees a substantial decrease in net imports due to large data center growth in NWPP and AZNM. In the east, regions that export heavily to regions such as SRVC-N see large impacts. Namely, RFCE, RFCM and RFCW also see their data center impacts more than double between ex post and forward looking periods. Our Mid-Range results indicate that the data center component of the wholesale price will make up over 10% of the wholesale price in 10 of our 23 regions, and more than 5% in 19 of our regions by 2028.

Given the convexity of the supply curve, it is not apparent from the average effects the hours during which data centers are most increasing the wholesale price. Research, such as [Norris et al. \(2025\)](#), has looked at how flexibility from data centers through curtailed compute during the highest demand hours can improve reliability of the grid and reduce the amount of required supply build-out to meet the proposed demand growth. However, wholesale price effects from data centers can occur at all points along the demand distribution. In Figure 6, we show the distribution of the total dollar amount of the price effect of data centers by the price decile of the hours. This measure accounts for both the size of the price effect and the amount of electricity consumed in that hour. In Panel A, we see that both the highest and lowest priced hours see outsized effects with smaller effects in the middle priced deciles. Importantly, around 75% of the effects from data centers are not concentrated in the highest price decile, meaning that curtailment strategies and mandatory flexibility can only reduce some of the wholesale price effect. The larger effects in the low priced hours are driven by the steep increase in the supply curve between cheap renewable sources and the thermal portion of the supply curve.

Figure 6: Price Effects by Hourly Price Decile, 2026–2028

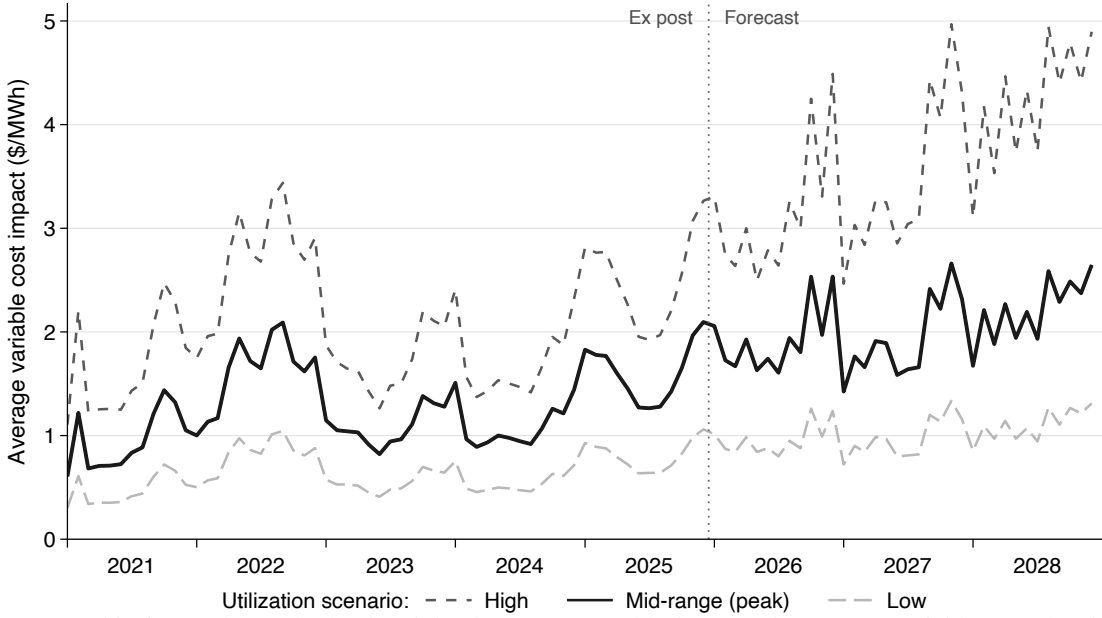


Notes: This figure shows how the increase in wholesale payments by non-data-center load is distributed across hours in 2026–2028, under the Mid-Range (Peak) utilization scenario. For each eGRID subregion and hour, the payment increase is the change in the wholesale price caused by data center demand times demand in the no-data-center counterfactual. Within each subregion and year, hours are ranked by their wholesale price in the no-data-center counterfactual. Panel A reports each price decile’s share of the total payment increase. Panel B splits the highest decile into hours between the 90th and 95th percentiles, the 96th–98th percentiles, and the 99th and 100th percentiles. Bars pool 2026–2028 and markers show individual years. Dashed lines mark the share each bin would carry if every hour contributed equally.

In non-restructured markets, vertically-integrated utilities operate under cost-of-service rate regulation. While the system lambda from our model describes the marginal cost for the marginal generator required to meet data center demand, the more relevant metric is the average variable cost of generation. In Figure 7 we show the change in average variable cost associated with data center expansion across the four data center consumption scenarios. Average variable costs are calculated as the total variable cost from thermal generation divided by total quantity of electricity consumed.³⁵

³⁵Note that this calculation provides a lower bound as it excludes variable costs from nuclear fuel and maintenance costs for renewable sources.

Figure 7: Realized and Projected Average Monthly AVC Impacts, 2021–2028



Notes: This figure shows the load-weighted average monthly increase in average variable costs by data center utilization scenario. AVC increases are computed as the difference between each eGRID subregion's hourly AVC under observed and counterfactual simulations and then averaged to the monthly level. Both series are weighted by the no-data-center demand. AVC impacts are in 2024 dollars per MWh.

5.2 Payments, Costs, and Rents

As our model assigns hourly dispatch instructions for specific generators, we are able to decompose the efficiency consequences of data centers. We use prices and quantities to calculate total competitive wholesale payments made to all generators and decompose the increase in payments from data center demand growth into changes in total variable production costs and competitive rents earned in the market.³⁶ 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} is total net generation in each subregion in hour t , including inframarginal generators, since all dispatched units are paid the wholesale price. Production costs are total variable costs (fuel, variable O&M, emissions permits) summed across all dispatched generators and hours. We define competitive rents equal to the difference between total wholesale payments and total production costs. Note that due to the convexity of the supply curve, competitive rents are possible even in a perfectly competitive market absent market power.

³⁶We follow past work, such as [Borenstein et al. \(2002\)](#), who similarly decompose these objects.

We further separate the competitive rents into inframarginal and marginal rents, where marginal is defined as the set of incremental generators who operate solely in the cases with data center demand.

Table 3: Data Center Impact on Payments, Production Costs, and Competitive Rents, Mid-Range (Peak) Utilization Scenario (Millions of 2024 Dollars), 2021–2028

	Ex-post			Cumulative	Forecast	Cumulative
	2021–2023	2024	2025	2021–2025	2026–2028	2021–2028
<i>National</i>						
Δ Payments	56,467	16,952	23,092	96,511	90,184	186,695
Δ Production Costs	24,371	7,787	11,512	43,670	44,221	87,891
Δ Total Rents	32,096	9,165	11,580	52,840	45,963	98,804
Δ Inframarginal	31,416	8,784	11,081	51,281	43,229	94,510
Δ Marginal	680	380	499	1,560	2,734	4,294
<i>Restructured</i>						
Δ Payments	36,575	10,895	15,109	62,578	59,805	122,383
Δ Production Costs	16,427	5,400	7,560	29,387	29,277	58,665
Δ Total Rents	20,148	5,495	7,548	33,191	30,528	63,718
Δ Inframarginal	19,600	5,223	7,199	32,022	28,621	60,643
Δ Marginal	548	272	349	1,169	1,907	3,076
<i>Non-Restructured</i>						
Δ Payments	19,893	6,057	7,983	33,932	30,379	64,311
Δ Production Costs	7,944	2,387	3,951	14,283	14,943	29,226
Δ Total Rents	11,948	3,670	4,032	19,650	15,436	35,085
Δ Inframarginal	11,816	3,561	3,881	19,259	14,608	33,867
Δ Marginal	132	108	150	391	828	1,218

Notes: Columns report the ex-post period in three blocks (2021-2023 combined, then 2024 and 2025 separately) and the forecast period as a 2026-2028 total. The two cumulative columns are the ex-post (2021-2025) and full-period (2021-2028) totals. All values are differences between the system with data centers and a no-data-center counterfactual under the Mid-Range (Peak) utilization scenario, in millions of 2024 dollars. ΔPayments are calculated as the incremental increase in each region’s wholesale price multiplied by its total net generation. ΔProduction Costs is the additional fuel, variable operating and maintenance, and emissions-permit costs of dispatched thermal generation needed to serve data center load. The assumes that inframarginal generation is held fixed between observed and counterfactual simulations. ΔRents is the difference between ΔPayments and ΔProduction Costs. Inframarginal rents are those attributable to generation that would have run in the counterfactual. Marginal rents are the surplus earned by the additional units dispatched in order to meet data center demand. We split these components by restructured and non-restructured markets as their interpretations vary between the two

In Table 3 we present the change in these three objects computed as the difference between dispatch in the Mid-Range Peak scenario and the no-data-center counterfactual. Data centers increased total payments to generators by \$96.5 billion over 2021-2025. Over our forecast period,

total payments are \$90.2 billion higher, representing a growth in average annual payments compared to the ex post period. Together, data centers are associated with a \$186.7 billion increase in total payments across all modeled periods.

The change in total payments aggregates two economically distinct components. First, it includes the cost of the electricity that data centers consume, valued at the prevailing market price ($\Delta ProductionCosts + \Delta MarginalRents$). The increase in electricity generation necessarily increases production costs, rising by \$87.9 billion over the full time horizon. Because the supply curve over the relevant portion is convex, marginal rents of \$4.3 billion accrue to the generators brought online over the full time horizon.

Second, and of relevance for the affordability debate surrounding data center growth, it includes the additional amount paid for electricity that would have been consumed regardless, arising solely because data center load moved the market to a higher price ($\Delta InframarginalRents$). This component constitutes a transfer from pre-existing consumers to generators. Nationally, inframarginal rents sum to \$94.5 billion over 2021–2028. Thus, roughly half of the increase in wholesale payments is borne by non-data center consumers.

We follow [Borenstein et al. \(2002\)](#) and further 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. Notably, we find a cumulative increase in payments from wholesale consumers to generators in restructured markets in our mid-peak scenario of \$62.6 billion from 2021 to 2025 that was driven by data center demand. The cumulative impact of data centers on competitive rents totaled over \$33 billion, the vast majority of which ($\approx 96.5\%$) went to inframarginal generators.

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 wholesale 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. While these numbers map less cleanly, the patterns in both cases remain strikingly consistent.

It is important to note that the incidence of such changes to payments and rents is outside the scope of our current research. This is complicated by the fact that many generators operate under power purchase agreements, under which they receive a contracted price for their generation. Similarly, many load serving entities may engage in PPAs or hedging for procurement. The nature of these contracts will ultimately dictate the timing and incidence of the wholesale market changes. These contracts may also delay the eventual pass through of these increases on to retail customers, a point we return to in the following section.

5.3 Emissions

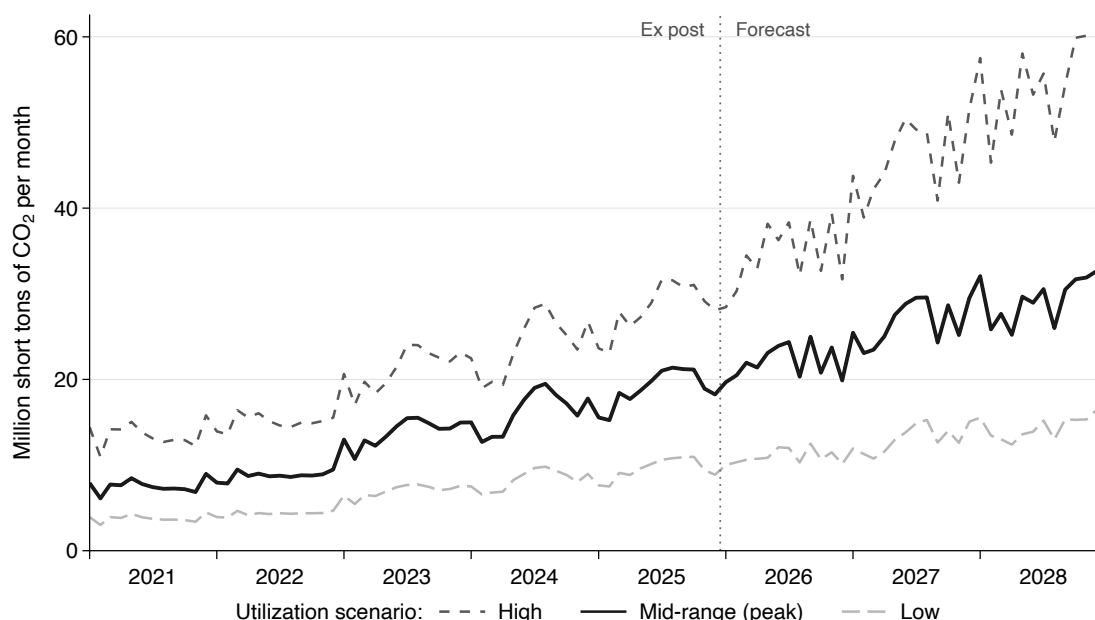
We next use the results of our dispatch to look at the change in grid-level emissions from the growth in data center electricity demand. As with price effects, the connected nature of the grid can lead to power plants far from the data center responding to the load growth, leading to emissions spillovers. Using plant specific emissions rates, we can estimate the incremental CO_2 emissions directly related to data center load growth. The main assumption, as with our price effect results, is that data center load growth is marginal, thus we attribute the incremental emissions between the data center and no data center case to data center demand.

Our results, shown in Figure 8, suggest that incremental emissions have been steadily increasing over our sample period as data center load continues to rise, suggesting that data center demand growth has generally outpaced the rate at which new renewable generating capacity has connected to the grid. By the end of 2025, data centers were generating between 8.9 and 28.1 million additional tons of CO_2 per month. For reference, Using a SCC of \$224³⁷ there translate in to around

³⁷We use the SCC estimate from [Rennert et al. \(2022\)](#) expressed in 2024 dollar per metric ton, consistent with the other dollar amounts in our paper

\$24 and \$75 billion in damages annually.³⁸

Figure 8: Realized and Projected Incremental Monthly CO₂ Emissions, 2021–2028



Notes: This figure shows the monthly incremental grid emissions by data center utilization scenario, in millions of short tons of CO₂. Emissions are computed from the unit-level heat rates and fuel carbon contents. These impacts only consider grid emissions; behind-the-meter generation from data centers or other facilities are not captured in these estimates.

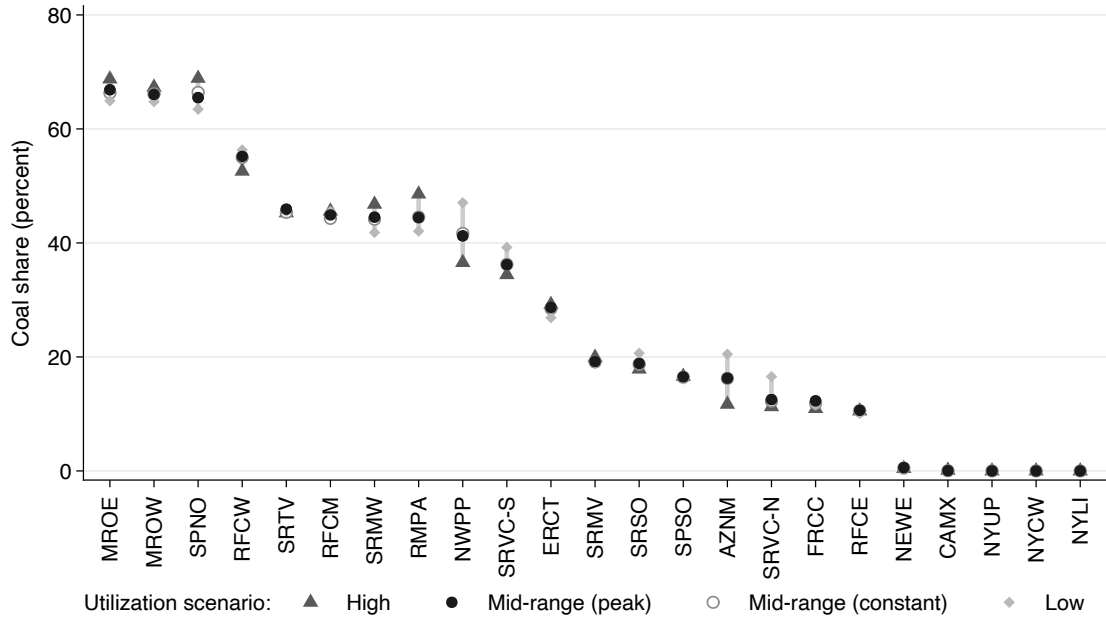
Looking forward, in the high-demand scenario, data center growth greatly increases emissions, resulting in between 45.3 to 60.1 million tons of CO₂ per month over in 2028. In the mid-scenario, emissions increases are still high, reaching over 25.2 million tons of CO₂ per month by 2028. It is important to note we only measure grid-generated emissions. Data centers may use backup generation or install on-site ‘behind-the-meter’ sources of generation. These sources of energy are often more heavily polluting than on-grid assets, and would be additional to our calculations.

Importantly, the heterogeneous generation mix across regions and time means that incremental demand is met by generation units of varying efficiencies and technologies. Thus, a data center’s impact on emissions depends on the marginal emissions rate in the specific region it is locating at that particular point in time. While heat rates and emissions rates can vary within technology, grids

³⁸As 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.

that rely on coal generation as the marginal source have a significantly higher marginal emissions rate. In Figure 9, we show the share of natural gas and coal marginal thermal generators used to meet data center demand across regions. In New England and California, the incremental demand is met almost entirely by natural-gas-fired generation, while in much of the upper Midwest, coal plays an important role.

Figure 9: Share of Incremental Load Met by Coal
By Region, 2024–2025



Notes: This figure shows the fuel composition of the incremental generators dispatched to meet data center demand under each utilization scenario from 2024–2025. We omit coke and petroleum plants given their shares are close to zero in all eGRID regions and scenarios.

6 Robustness and Extensions

This section examines the wholesale market effects under a number of alternative assumptions and scenarios. In Section 6.1 we consider an alternative scenario where trade flows are fixed. In Section 6.2 we show our results are robust to our assumptions on the evolution of electricity supply by considering alternative build-out and retirements scenarios as well as allowing new supply additions to endogenously respond to data center entry. Finally, in Section 6.4 we consider market impacts if data center demand could flexibly move to regions with the lowest cost of supplying

electricity.

6.1 No Trade Flows

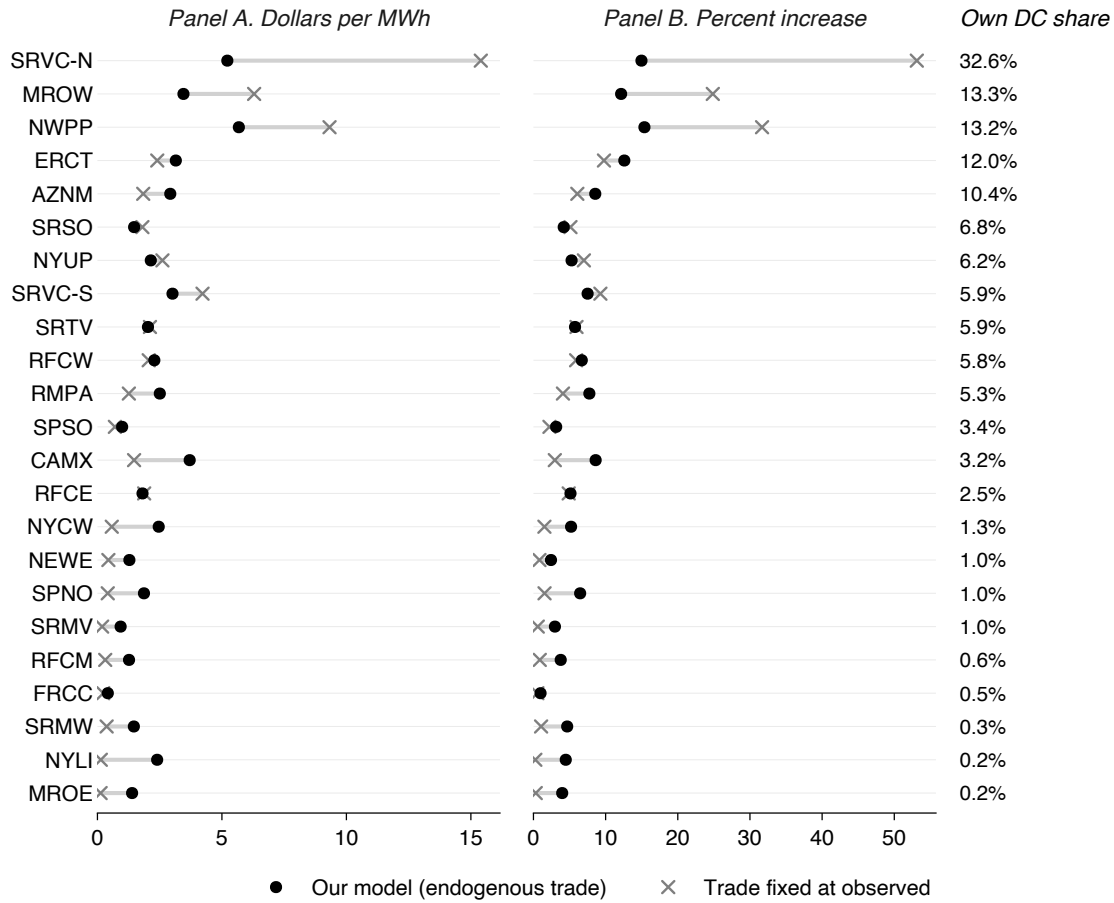
In Section 4.3, we showed that our model builds upon a class of dispatch models which do not account for cross region transmission and replicates observed trade flows accurately. In this exercise, we document the importance of accounting for trade flows in the dispatch model. To do this, we remove the trading mechanism between regions and use the hourly observed net thermal generation as the implied residual demand in the region as in [Hausman \(2025\)](#); [Ham et al. \(2025\)](#). This implicitly fixes the trade from other regions at the observed level. For the no data center counterfactual, the hourly regional demand is lowered by the assumed data center usage and the model is re-solved, forcing all data center demand to be met by within region generators.³⁹

In Appendix Table A2, we show that both versions of the dispatch model recover similar nationwide average price effects, percentage changes, and emissions outcomes from data centers during the ex post period. Due to the lack of trade, the no-trade versions has more hours with unserved energy where the regional demand is greater than supply.

However, in Figure 10, we plot the difference in the regional average hourly price effect from data center growth over 2024–2025 between the two model versions. As expected, regions that experience large demand growth from data centers, such as Virginia (SRVC-N), the upper midwest (MROW), and the northwest (NWPP), experience much larger price effects in the version where trade is held fixed as more expensive local generation is required to meet the higher demand. Also, a region like California, CAMX, who has trade partners experiencing large demand increases has their price effect understated in the no-trade version since the loss of foregone trade is not accurately captured.

³⁹We continue to account for unit ramping constraints so that we can isolate the importance of the trade flows in our impacts.

Figure 10: Difference in Wholesale Price Effects Between the Trade Model and a Model Fixing Trade at Observed Levels



Notes: This figure shows the difference in average hourly wholesale price effects of data center growth under our mid scenario between our baseline model with trade and a version where trade flows are held fixed as in [Hausman \(2025\)](#).

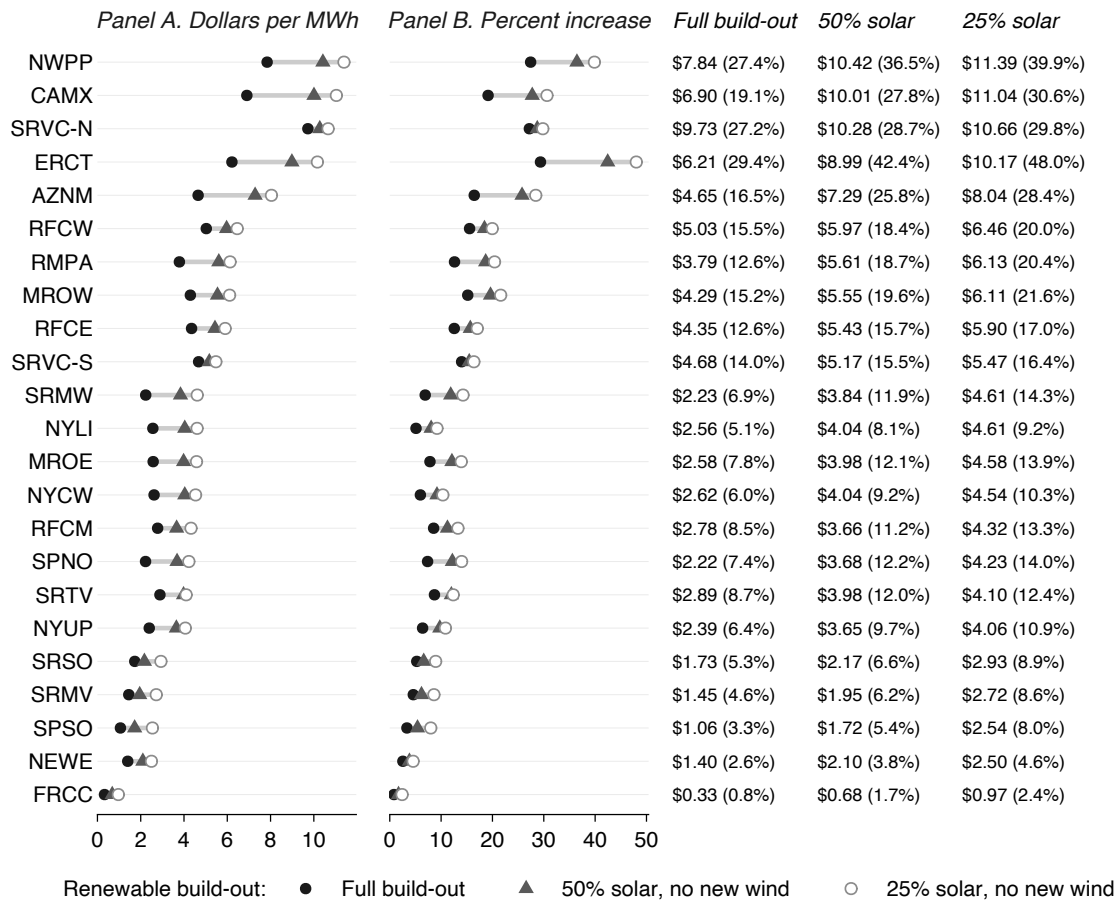
6.2 Alternative Evolution of Electricity Supply

Next, 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 needed to meet the increase in demand. The magnitude and timing of new generation additions to the grid can be important for determining wholesale market effects.

6.2.1 Alternative Solar and Wind Build-Out

We estimate the price increase 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 alternative supply scenarios. The black dots show the increase in total prices from data centers under the baseline Mid-Range (Peak) scenario. The additional points represent the additional price increases that would result from a slower build renewables build path of either a 25% reduction or 50% reduction in proposed solar and no new wind plants.

Figure 11: Projected Percentage Increase in Wholesale Prices
Under Alternative Renewable Build-Out, 2028



Notes: This figure shows the increase in wholesale prices attributable to data centers under our three renewable supply build-out paths. The black dot represents the price increase attributable to data centers under the full build out scenario. The other points represent the additional price increases that would come from reducing renewable build out.

These alternatives show the mitigating effects that the planned renewable build out will have on the effects imposed by data centers, if the data center build out realizes. With fewer additional low-marginal-cost renewable resources available, incremental load is more likely to be served by higher-marginal-cost thermal generation. This is particularly noticeable in the solar-rich parts of the west such as the southwest (AZNM), Texas (ERCT), California (CAMX), and the northwest (NWPP). 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. Appendix Figure A6 details the emissions impacts of the alternative renewable build-outs.

Our supply forecast also assumes that power plants retire at their announced date. Recently, there has been political pressure to keep coal power plants online longer, extending their operational life. In Appendix Figures A7 and A8 we update our supply forecast to reverse planned retirements.⁴⁰ The results presented in Appendix Figure A7 show the difference in price effects across the four data center demand scenarios. Price effects are mostly negligible; our Mid-Range scenarios imply a maximum 1.84 percentage point difference in average monthly wholesale prices. The difference in emissions, presented in Figure A8, implies an even smaller divergence of less than -0.049 percentage points. Taken together, these results highlight an important result. Given coal capacity makes up the vast majority of retirements, current policies that force coal plants to remain online may not produce the intended results that they champion due to the uneconomic nature of coal.

6.2.2 Endogenous Entry

One 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 our forward-looking time horizon would do so even if there was no demand

⁴⁰We aggregate this set of results to the yearly level because the non-linearities introduced by this counterfactual make the monthly figures less-interpretable.

from data centers. We believe this assumption is reasonable in the short run. Given the lengthy interconnection queue, which even in the quickest regions can take 5-7 years ([Lawrence Berkeley National Laboratory and GridTracker 2026](#)), it is unlikely that observed thermal generation in our panel is a result of the boom in AI demand which arose at the beginning of 2023, implying that most of the projects were likely already under development before data center demand was projected to be as large as current projections. However, some renewable generation and battery storage is able to connect faster driven by faster permitting timelines.

In the long run, it is plausible that new supply would enter specifically to address data center demand. Policies requiring data centers to “bring-your-own-generation” are under discussion in many jurisdictions, requiring data centers to pay to construct or contract with new sources of generation in order to be allowed to connect to the grid.⁴¹ Including these endogenous 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 scenario and overstate the wholesale price effects caused by data centers. To address this concern, we consider the case where 25% of new supply is marginal to data center demand and would not have entered had there been no data center demand.⁴²

Appendix Figures [A9](#) and [A10](#) 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 2028, our main result overstates the effect of data centers by, at most, 2.44 percentage points relative to the endogenous supply case. For emissions, the difference starts in January 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. Furthermore, it is worth noting that our decision to assume that 25% of new supply is endogenous to data center demand is assuming a fairly large share of marginal entrants and we therefore believe

⁴¹There was not, to the best of our knowledge, any binding bring-your-own-generation legislation applicable during the ex post portion of our study.

⁴²We implement this by derating the capacity of all new entrants and observed capacity additions to existing generation by 25% only in the ‘no data center’ counterfactual.

this represents a conservative test of the exogenous supply assumption.

6.3 Demand Elasticity and Non-Data-Center-Demand Crowd-Out

In our baseline counterfactual, we remove data center load and hold all other consumption at observed levels. If data centers increase prices for existing customers, they may consume less than they otherwise would have. Equivalently, in a counterfactual world without data centers, these same customers would consume more power and raise the counterfactual price. Ignoring this response would mean that our model overstates data center impacts. We bound the size of that bias for our Mid-Range (Peak) scenario by letting non-data center demand respond to data center price effects.

Few customers face real-time wholesale electricity prices, and estimated real-time price elasticities are approximately zero (Fabra et al. 2021). In addition, following the results in Ito (2014), we model the response through average retail prices, since consumers tend to respond less to marginal prices. Leaving our hourly dispatch unchanged, we consider the effects of applying a constant elasticity to the percentage change in each subregion’s average annual retail price.

Formally, the residual demand in each subregion r and hour t in our no-data-center counterfactual becomes:

$$\tilde{d}_{rt} = d_{rt} + \left[\left(1 - \frac{\beta_{ry}}{P_{ry}} \right)^\varepsilon - 1 \right] D_{rt}, \quad (12)$$

where d_{rt} is the counterfactual no-data-center residual demand from Equation 3 and D_{rt} is the hourly demand from existing, non-data center customers in each subregion. The average wholesale price impact of data centers in each subregion-year y —the load-weighted average difference between the price with and without data centers—is β_{ry} , while the average annual retail price in each subregion is P_{ry} . This is computed from EIA-861 as the revenue per MWh across all customer classes.⁴³ Finally, ε represents the price elasticity of demand. Notice that Equation 12 implies full

⁴³This includes residential, commercial, industrial, and transportation customers. We use the 2024 data for 2025.

pass-through of the wholesale price effect. This assumption is made both for simplicity and to maintain a conservative bound in this exercise. We assume that added demand is spread across hours in proportion to non-data center demand. In addition to the change in the price response, we define the crowd-out share, θ_y :

$$\theta_y = \frac{\sum_r \left[\left(1 - \frac{\beta_{ry}}{P_{ry}} \right)^\varepsilon - 1 \right] \sum_{t \in y} D_{rt}}{\sum_r \sum_{t \in y} DC_{rt}}, \quad (13)$$

where DC_{rt} is hourly data center consumption in each subregion. This is simply the added non-data center consumption as a share of the data center consumption over the period.

We consider three values of ε to capture short, medium, and long-run time-horizons within which consumers can adjust their behavior. We set $\varepsilon = \{-0.1, -0.3, -1\}$, following the estimates from the literature.⁴⁴ The relevant time horizon is not obvious here. In our counterfactual we remove all data center load, which includes approximately 134 TWh attributable to data centers that were already operational in 2021. On the other hand, data center load grew more rapidly during our period of study from 2021 to 2025. A lagged response to that growth would likely be smaller than a contemporaneous one.

Given counterfactual load, we solve the same dispatch model from our main results in every hour. The counterfactual demand response increases the associated price, shrinking β_{ry} and thus $\left(1 - \frac{\beta_{ry}}{P_{ry}} \right)^\varepsilon - 1$. Therefore, we solve for the fixed point around the unchanged dispatch model in order to recover the demand response under each counterfactual elasticity.

Table 4 reports the nationwide results. In each specification, we find that including elasticity in demand reduces the data center effects roughly in proportion to the elasticity used. Under our central case of $\varepsilon = -0.3$, a crowd-out share of $\theta_r = 10.5\%$ implies existing customers would have

The forecast scales non-DC load in the with-DC scenario by $\left(1 - \frac{\beta_{ry}}{P_{ry}} \right)^\varepsilon$, with P_r escalated 2.5% a year.

⁴⁴Burke and Abayasekara (2018) estimate a same-year elasticity of -0.1 and Deryugina et al. (2020) estimate ε at -0.09 within six months following a change in prices. Deryugina et al. (2020) also inform our central case of -0.3. They estimate a response of -0.27 two to three years after a price shock, and a recent, extensive meta analysis from Kudela et al. (2026) provide a bias-corrected long-run estimate of -0.38. For an upper bound, we use the long-run estimate of (typically more price-responsive) industrial customers from Burke and Abayasekara (2018), where they find values near -1.

consumed an additional 115 TWh of electricity in the no-data center counterfactual. We find that nationwide prices would have been 0.7 pp and 1.3 pp lower in the ex post and forecast central cases, respectively. Under our upper bound of $\varepsilon = -0.1$ and full pass-through, we find that the majority of the price effect persists. There are two other results worth noting. First, due to the convexity of supply, one concern might be that the share of inframarginal rents we compute relative to total payments could be overstated under our baseline counterfactual. In the ex post results, this share moves from 0.531 to 0.527 moving from $\varepsilon = 0$ to -1 . In the forecast, the share moves slightly more from 0.479 in our baseline to 0.471 in the central case, and 0.458 in the upper bound, largely assuaging those concerns. In the forecast period, the move from the baseline to central case implied a lower emissions reduction than in the ex post period.

Table 4: Data Center Effects When Non-Data-Center Demand Responds to Prices, Mid-Range (Peak), 2021–2028

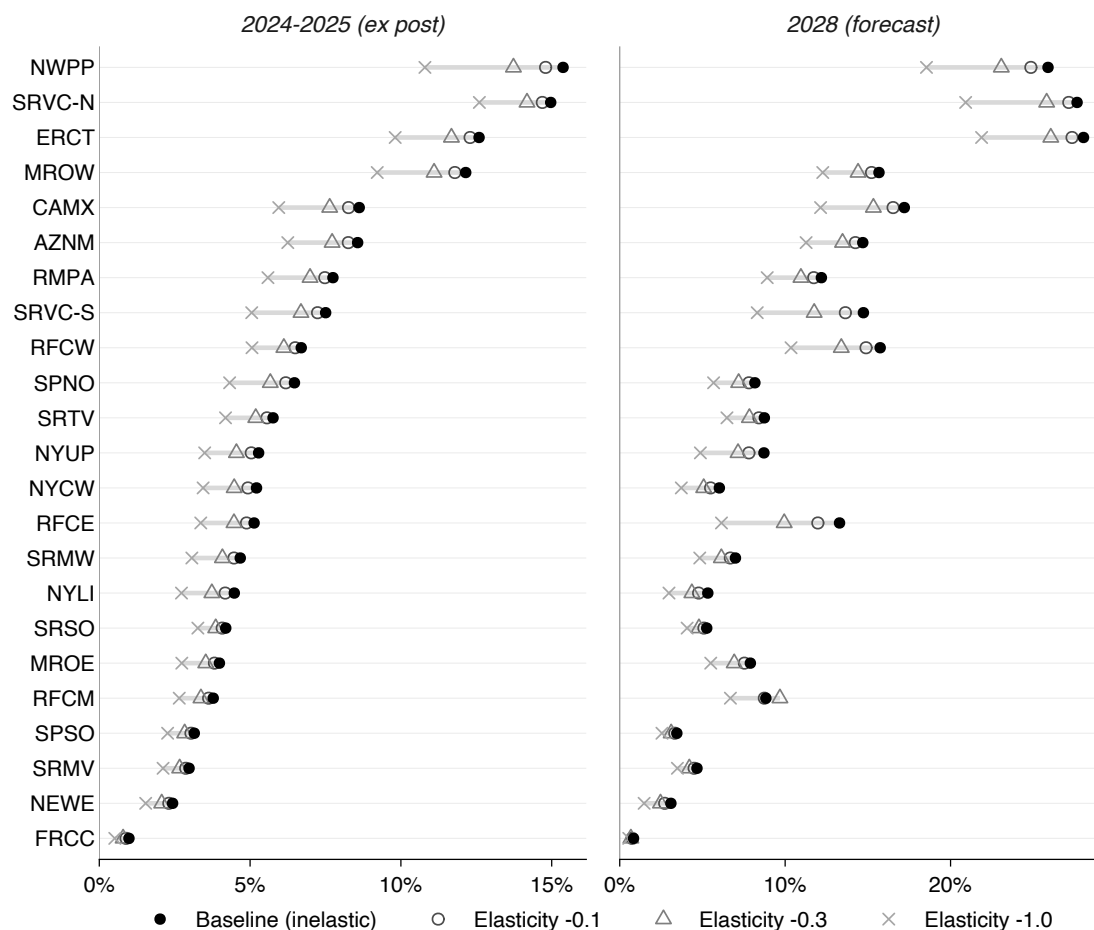
	Crowd-out	National price effect			Change over the period			
	share	Final	Final	Period	Cost	Pay-	Infra-	CO ₂
	(%)	year	year	mean	(\$B)	ments	marginal	(Mt)
		(\$/MWh)	(%)	(%)		(\$B)	(\$B)	
<i>A. Ex post, 2021–2025 (final year 2025)</i>								
$\varepsilon = 0$ (baseline)	0.0	2.66	6.9	5.9	43.7	96.5	51.3	784.0
$\varepsilon = -0.1$	3.8	2.58	6.6	5.6	41.8	92.2	49.0	754.2
$\varepsilon = -0.3$	10.5	2.42	6.2	5.1	38.6	84.7	45.0	702.4
$\varepsilon = -1.0$	27.1	1.96	5.0	4.0	30.8	66.7	35.2	571.7
<i>B. Forecast, 2026–2028 (final year 2028)</i>								
$\varepsilon = 0$ (baseline)	0.0	4.01	12.3	10.0	44.2	90.2	43.2	937.4
$\varepsilon = -0.1$	2.5	3.84	11.8	9.6	43.0	86.9	41.4	913.2
$\varepsilon = -0.3$	6.9	3.56	11.0	8.9	40.9	81.4	38.3	870.3
$\varepsilon = -1.0$	18.7	2.85	8.8	7.2	35.3	68.1	31.2	756.3

Notes: Each row re-solves the model with non-data-center demand that responds to the data centers' effect on its average retail price with constant elasticity ε .

Figure 12 displays the effects of price-responsive demand on wholesale prices by eGRID subregion for the ex post and forecast periods. The results largely preserve the geographic patterns we find in the main results. A notable exception lies in the comparison between Northwest Power Pool and Northern Virginia in the ex post period. In the central case ($\varepsilon = -0.3$), the effect in

the Northwest falls from 15.4% to 13.7% while Northern Virginia's results only fall from 15% to 14.2%.

Figure 12: Wholesale Price Effects by Subregion Under Price-Responsive Non-Data-Center Demand



Notes: This figure shows the effect of data centers on the wholesale price in each eGRID subregion from 2024–2025 and in 2028 under our baseline elasticity assumption and three other estimates that allow for price-responsive non-data-center demand.

Taken together, these results imply a modest upward bias, at most under a set of plausible demand elasticities. This will hold especially true if the assumption of full pass-through from wholesale to retail rates fails. Under all of our cases, we find similar effects on prices and emissions, and the price effects remain concentrated in the same subregions.

6.4 Geographic Compute Flexibility as Virtual Transmission

Finally, we explore the extent to which data center load flexibility can mitigate the wholesale market effects discussed in Section 5. 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.

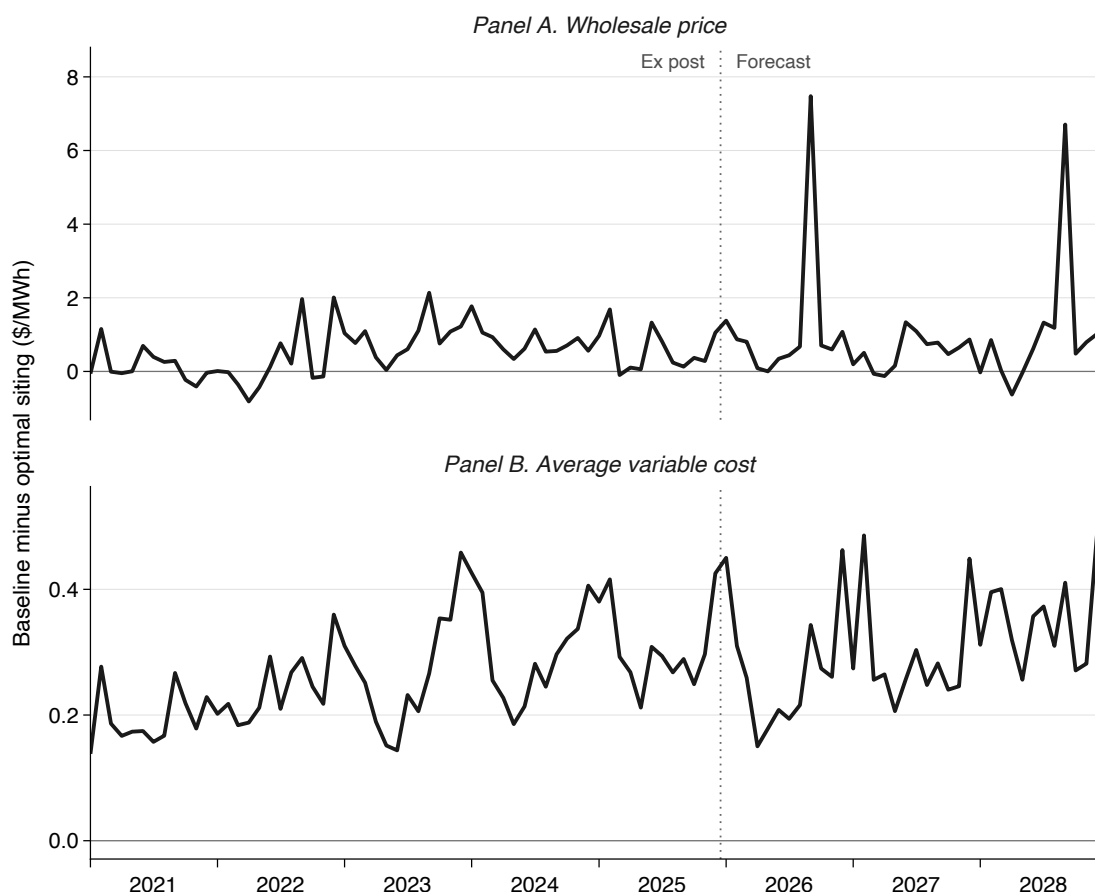
To illustrate the potential magnitude of this margin, we compare observed data center siting against the geographic allocation of compute chosen by a social planner minimizing short-run electricity operating costs.⁴⁵ This counterfactual abstracts from latency, construction costs, labor availability, interconnection constraints, and other factors that shape actual siting decisions. We solve a national dispatch problem, allocating excess thermal capacity that remains *idle* in the no-data-center scenario to serve compute. The resulting hourly generation pattern identifies where compute would be served if short-run (hourly) electricity operating cost were the sole objective and there was sufficient data center capacity in each region.

Figure 13 shows the magnitude of changes to market outcomes from allowing data center demand to be geographically flexible. From 2021 through 2025, the monthly average wholesale electricity price would be up to \$2 lower under the hourly cost minimization strategy though with considerable seasonal volatility. This wedge represents approximately 1%–3% of total system generation costs. Looking forwards, these potential price reductions reach up to \$8 during particularly

⁴⁵For full details on how we solve the social planner’s problem, see Appendix H.

tight supply conditions as overall data center demand increases.

Figure 13: Difference in Market Outcomes Between Observed and Spatially Optimal Compute

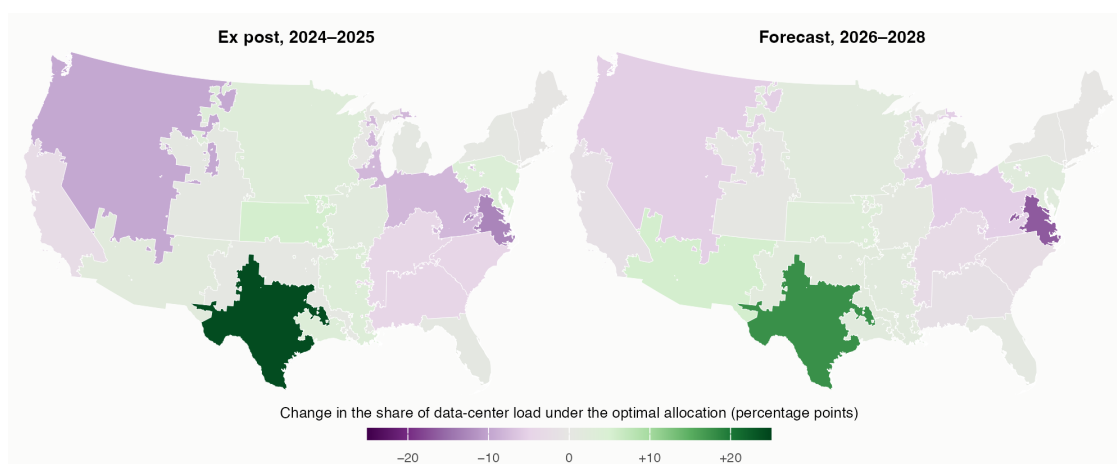


Notes: These figures show the difference in average national monthly electricity price and monthly average variable costs between observed data center compute under the Mid-Range (Peak) utilization scenario and the cost-minimizing spatial allocation of compute which allows for data center demand to flexibly move to the lowest cost generation sources available. The cost-minimizing spatial allocation is still subject to network constraints, such as interface and import capacities.

Figure 14 shows that the cost minimizing geographic allocation of compute differs dramatically from the observed patterns. Over 2024–2025, ERCOT accounts for 19.9 percent of realized data center compute in the data, whereas the cost minimizing allocation would allocate approximately 44 percent there—a gap of roughly 24 percentage points. By contrast, Virginia (SRVC-N) accounts for 14.6% of realized compute over this period, while the cost-minimizing allocation assigns it only 1.7% due to the high pre-existing wholesale prices. Under the planner allocation, just five regions—Texas (ERCT), the upper midwest (MROW), Arizona/New Mexico (AZNM), the northern Mid-Atlantic (RFCE), and SPP-North (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 slightly less concentrated. Under the forecasted system, 75 percent of compute in the cost-minimizing allocation is served by six regions. 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 14: 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 Retail Price Effects

Section 5 shows that data center electricity demand has had a significant impact on wholesale electricity prices and that this impact could be even larger in the near future. However, much of the public debate centers on the retail rates that residential customers actually pay. Industry groups and a number of recent papers and reports have argued that additional load growth from data centers will actually cause retail prices to decrease due to the fixed-cost-spreading effect. However, as we show below, the benefits from the spreading of fixed costs will be offset by the increase in the short-run marginal cost of producing electricity discussed in Section 5.

The effect of new data center load on customers' bills is theoretically ambiguous. Under

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. This fixed-costs-spreading could lower the burden borne by existing customers on the grid.⁴⁶ This outcome is especially relevant if data centers locate in regions where there is pre-existing spare transmission and distribution capacity or flexibly decrease their consumption during periods of peak load (Norris et al. 2025).

We present a simple framework in Appendix G to quantify the relative magnitudes of the cost increases from higher variable costs and the cost decreases from the spreading of fixed costs. To do so, we invoke Assumption 1, assuming that additional electricity demand will only increase the total system costs through the effect on short-run marginal costs modeled in Section 5.

Assumption 1. Additional electricity demand only increases total system costs through the increase in short-run marginal generation costs. All fixed system costs remain fixed.

This assumption is unlikely to hold in reality. 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, “fixed costs” will not actually be fixed, as large system wide infrastructure costs are needed to connect additional load.⁴⁷ We therefore consider the analysis in this section as likely to understate the retail price effects from data center load, showing a scenario where additional load is added to the system without requiring any system upgrades.⁴⁸

If fixed costs truly remain fixed, we can determine the share of the benefits from fixed cost spreading that are offset by the increase in the short-run marginal cost of producing electricity. We

⁴⁶Related 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).

⁴⁷Many developers and hyperscalers have pledged to pay for infrastructure upgrades, but regulatory and technical obstacles make implementing a clear system for them to do so challenging. However, in the case that they do cover all incremental fixed costs, then our assumption holds.

⁴⁸We also make the simplifying assumption that costs are recovered uniformly across all customer classes. In reality, commercial and industrial customers often face lower electricity rates than residential customers. In the limit, where data centers 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.

call this measurement the offset ratio (OR) and in Appendix G show that it can be described by

$$\text{Offset Ratio} = \left(\frac{1-s}{s} \right) \cdot \frac{1+d}{\eta} \quad (14)$$

where d is the percent increase in electricity demand caused by data centers, s is the share of the retail bill that goes towards covering system fixed costs (the fixed cost share), and η is the electricity supply price elasticity.⁴⁹ For higher levels of OR, the increase in wholesale prices offset more of the benefits of fixed-cost spreading. If $OR > 1$, then the increase in wholesale prices more than fully offsets the fixed-cost-spreading benefits and it will be impossible for additional load to lower average costs, even if Assumption 1 holds and system fixed costs remain completely unchanged.

Equation 14 shows that, for an incremental increase in demand (small d), the relative magnitudes of the fixed-cost-spreading effect and wholesale price effects are entirely determined by the supply elasticity (η) and the fixed cost share (s). If most of the system costs are fixed costs (s is large) then the benefits of fixed cost spreading are more likely to dominate and wholesale price increases are unlikely to offset many of these benefits leading to a small OR . On the other hand, if electricity supply is very inelastic (small η), supplying additional electricity will lead to large wholesale price increases, and the wholesale price effect is likely to be more important, increasing OR . Finally, for non-incremental changes in demand, the percent increase in electricity demanded, d , will matter as well.

Equation 14 also implies that there is a break even fixed cost share, s , where the wholesale price increases exactly offset the fixed-cost-spreading benefits and additional load leaves average variable costs unchanged, putting neither upward nor downward pressure on retail rates. We solve

⁴⁹Here the relevant “price” will depend on the regulatory jurisdiction and market environment. In areas with wholesale markets, the relevant price is simply the wholesale price. In areas without wholesale market, average variable generation costs are likely a better approximation for the costs that get passed through to retail bills.

for this critical fixed cost share, s^* , by setting $OR = 1$ and find that

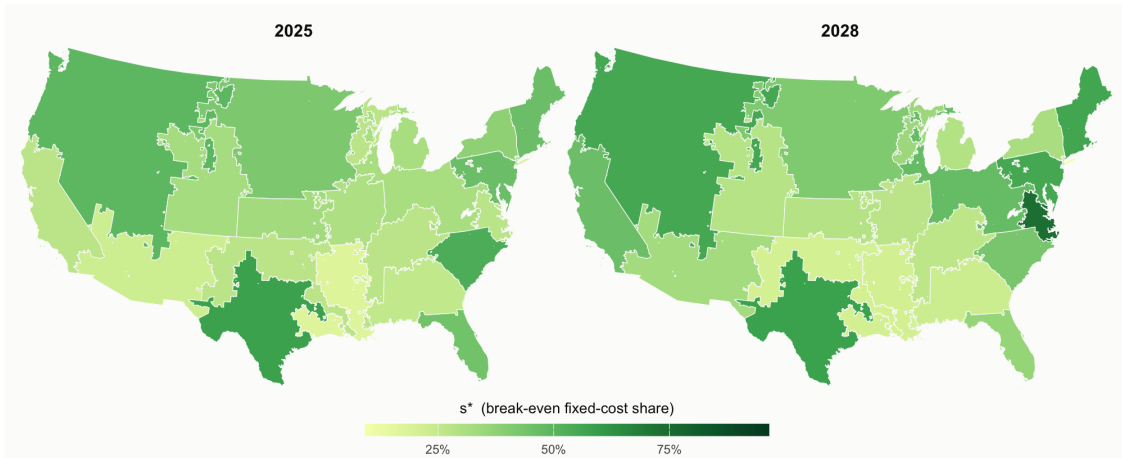
$$s^* = \frac{1+d}{1+d+\eta}. \quad (15)$$

Load growth can therefore only put downward pressure on retail rates if $s > s^*$.

Using our model of wholesale electricity markets, we can calculate the supply elasticity, η , and the share of electricity demand from data centers d for each region. We calculate η_r for region r by perturbing total hourly electricity demand by 5% in region r using the mid-peak demand scenario. Solving the model with the perturbed values of demand and our baseline demand in region r allows us to calculate the change in wholesale prices in region r and therefore η_r .

We map the resulting η_r for 2025 and 2028 in Figure 15. The darker regions indicate areas where the variable cost increases are larger, making it necessary for there to be a higher fixed cost share in order for load growth to put downward pressure on retail bills. There is considerable heterogeneity over time and between regions.

Figure 15: Threshold Fixed Cost Share Required for Rate Relief:
Local Demand Shock (Excludes Spillovers)

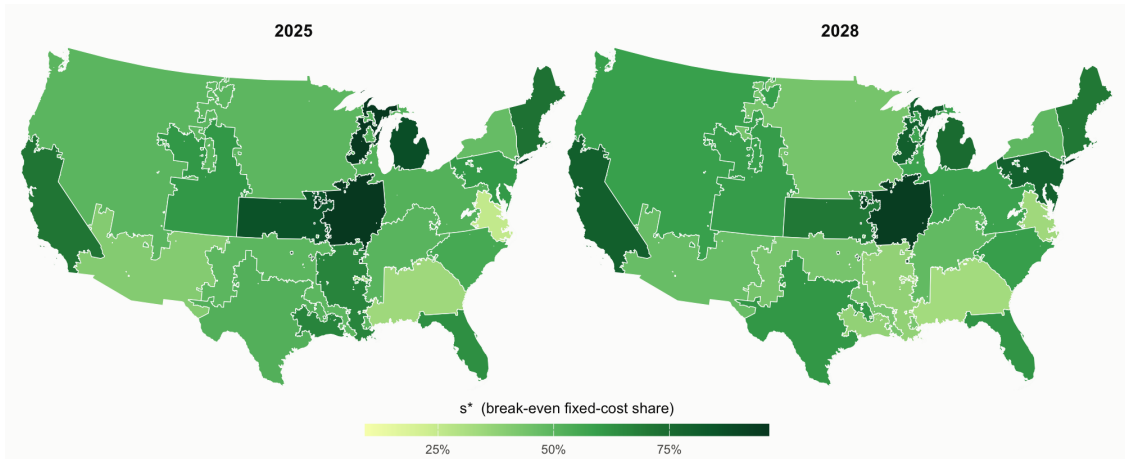


Notes: This map shows the minimum fixed cost share such that it would be possible for new load to lower residential bills, assuming that new load does not increase total fixed costs at all. The analysis uses load-weighted average wholesale prices as the relevant measure of variable cost pass through to end-use customers. The elasticity η_r is calculated for each region one at a time, and therefore does not account for spillovers from load growth in other areas.

Figure 15 shows the s^* from calculating η_r using the change in wholesale prices from only

changing demand in region r . However, changing demand in region r not only impacts electricity prices locally, but also has spillover effects on electricity prices in other regions through trade. We therefore calculate an alternative version of η_r that captures these spillover effects from data center growth. Here we calculate η_r as the percent increase in demand from datacenter in region r under the mid-peak demand assumption divided by the difference in wholesale prices in region r between the “mid-peak” and “no-data-center” demand scenarios. The break-even fixed cost share, s^* , using this alternative definition of η_r is shown in Figure 16. Overall, the map is considerably darker than the local estimates, suggesting that accounting for spillovers makes the variable cost increases relatively more important, putting upward pressure on retail rates.

Figure 16: Threshold Fixed Cost Share Required for Rate Relief:
Global Demand Shock (Includes Spillovers)



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 load-weighted average wholesale prices as the relevant measure of variable cost pass through to end-use customers. The elasticity η_r is calculated by taking the difference in the wholesale electricity price between the “mid-peak” and “no-data-center” demand scenarios.

Notably, many of the regions that are the darkest do not host many data centers. For example, the region that makes up much of central MISO in Missouri and southern Illinois (SRMW) would require fixed costs to be nearly all of the retail bill in order for fixed-cost-spreading to outweigh the variable cost increases. Appendix Figures A3 and A4 show that this region is home to very little data center capacity, but is near a lot of data center capacity in the upper midwest (MROW) and western PJM (RFCW). The high export capacity means that this significant neighboring data

center capacity will drive up wholesale prices in SRMW, but there is almost no local load for fixed cost spreading.

In order to determine the ultimate effects on retail bills, we need to calibrate the share of the retail bill that goes towards covering fixed costs, s . Measuring s is a non-trivial task that would require in-depth accounting of utility costs on a utility-by-utility basis, so we instead approximate it with one of two measures (with more detail in Appendix G.1). In wholesale markets, we follow the approach taken by [Borenstein and Bushnell \(2022\)](#) and use the average wholesale price to reflect the short-run variable costs. We then measure the variable cost share as the load-weighted average wholesale price divided by the average retail price calculated from EIA-861 data and the fixed cost share as one minus the variable cost share.⁵⁰ In areas without wholesale markets, we instead need to rely on costs reported by investor-owned utilities.⁵¹ Here we calculate the variable cost share as the reported spending on fuel and purchased power divided by total operating revenue. The fixed cost share is then calculated as one minus the variable cost share.⁵²

In Figures 17 and 18, we use these approximations of s and map the offset ratio, which measures the share of the fixed-cost-spreading benefits that are offset by variable price increases. Figure 17 shows this measure using only the local demand shocks and therefore ignoring spillovers. This figure suggests that in many parts of the country it is entirely possible that the benefits of fixed cost spreading could outweigh the increased variable cost from local data centers. In other words, if local load can be added to the grid without increasing fixed costs at all, this load-growth could put downward pressure on electricity bills. The dark blue regions in the southwest and southeast suggest that these regions could locally benefit the most from load growth. The ERCOT grid in Texas is the only region with an offset ratio larger than one in both 2025 and 2028, suggesting that data center load growth would put upward pressure on retail rates, even if there was no additional

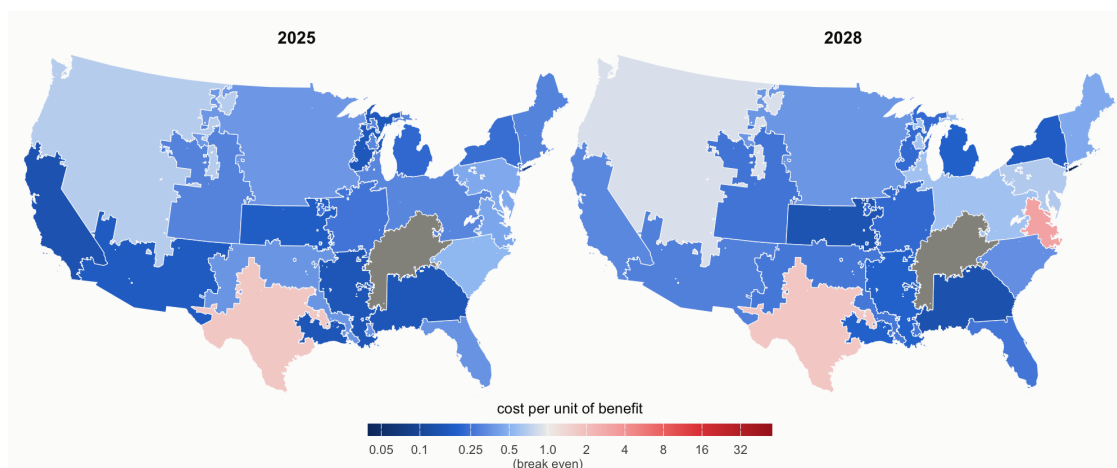
⁵⁰This assumes that all costs that are not short-run marginal costs are fixed and can be spread over more MWh. In reality, parts of the retail bill (like taxes) are neither short-run marginal costs of electricity production or fixed poles and wires costs that can be spread over more MWh.

⁵¹All investor-owned utilities file financial reports with FERC reported on Form 1. However, these reports do not include any reporting from publicly owned utilities or cooperatives. We access FERC Form 1 data through the Public Utility Data Liberation project (<https://catalyst.coop/pudl/>).

⁵²The region covering Tennessee that is largely served by the TVA has neither wholesale markets nor a large share of investor owned utilities so is excluded from this section of the analysis.

infrastructure spending required to connect data centers to the grid.

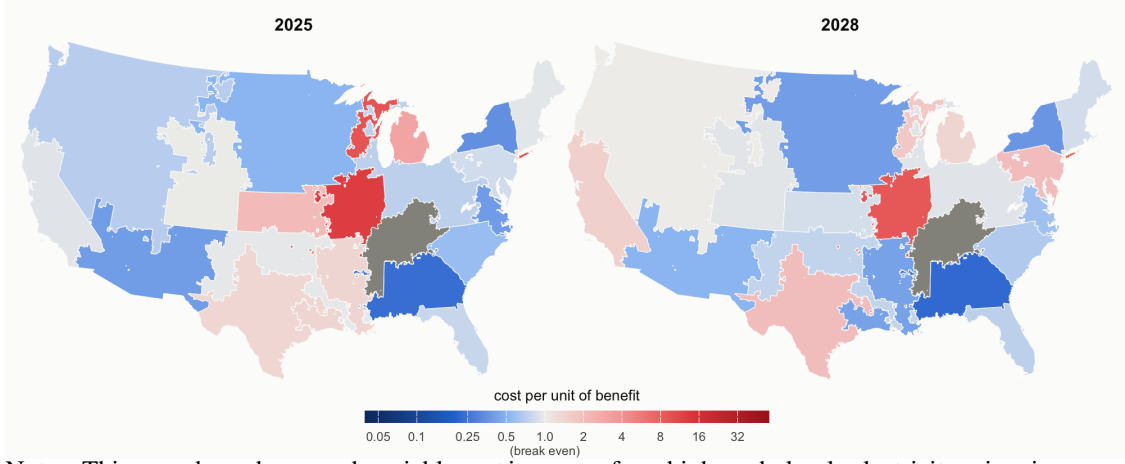
Figure 17: Offset Ratio: Local Demand Shock (Excludes Spillovers)



Notes: This map shows how much variable cost increases from higher wholesale electricity prices increases retail rates for every \$1 reduction in retail rates from fixed cost spreading. The analysis assumes that new load does not increase total fixed costs at all and uses load-weighted average wholesale prices as the relevant measure of variable cost pass through to end-use customers. Values greater than one indicate that the variable cost effects are larger than the fixed cost benefits, and new load cannot lower retail rates, even if there are no increases in system costs. The elasticity, η_r , is calculated by taking the difference in the wholesale electricity price between the “mid-peak” and “no-data-center” demand scenarios. The fixed cost share, s , is calibrated as either one minus the share of the wholesale rate divided by the retail rate (in wholesale markets) or as the share of total costs comprised of capital expenditure costs.

Accounting for spillover effects makes the outlook for retail rates less promising. Figure 18 shows the offset ratio using the global demand shock and therefore accounting for wholesale pricing spillovers. Considerably more regions than in Figure 17 now have offset ratios greater than 1, showing up as red, suggesting that accounting for spillovers is important for analyzing the effects data centers have on retail rates and makes the upward pressure on rates from wholesale price increases relatively more important. The offset ratios exceeding one mean that it is impossible for the data center load growth to lower retail rates.

Figure 18: Offset Ratio: Global Demand Shock (Includes Spillovers)



Notes: This map shows how much variable cost increases from higher wholesale electricity prices increases retail rates for every \$1 reduction in retail rates from fixed cost spreading. The analysis assumes that new load does not increase total fixed costs at all and uses load-weighted average wholesale prices as the relevant measure of variable cost pass through to end-use customers. Values greater than one indicate that the variable cost effects are larger than the fixed cost benefits, and new load cannot lower retail rates, even if there are no increases in system costs. The elasticity, η_r , is calculated by taking the difference in the wholesale electricity price between the “mid-peak” and “no-data-center” demand scenarios. The fixed cost share, s , is calibrated as either one minus the share of the wholesale rate divided by the retail rate (in wholesale markets) or as the share of total costs comprised of capital expenditure costs.

These results highlight how important accounting for the interconnected nature of the grid is for determining the effects of data centers on electricity rates. Transmission between regions allows lower cost power generation to meet demand elsewhere in the country when it is available, allowing the total cost of producing electricity to be lower. However, it also means that electricity markets are inherently highly interconnected and changes to either electricity supply or demand in one location will impact markets all across the country.

8 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. 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 Northwest. 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 the model implies meaningful increases in wholesale market pressure across all scenarios.

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 an 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 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.
- **and James Bushnell**, “The US Electricity Industry After 20 Years of Restructuring,” *Annual Review of Economics*, August 2015, 7 (1), 437–463.
- **, 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.
- Burke, Paul J and Ashani Abayasekara**, “The price elasticity of electricity demand in the United States: A three-dimensional analysis,” *The Energy Journal*, 2018, 39 (2), 123–146.
- Bushnell, James**, “A mixed complementarity model of hydrothermal electricity competition in the western United States,” *Operations research*, 2003, 51 (1), 80–93.
- Cicala, Steve**, “Imperfect markets versus imperfect regulation in US electricity generation,” *American Economic Review*, 2022, 112 (2), 409–441.
- Cipher Mining Inc.**, “Cipher Mining Announces September 2025 Operational Update,” Press release October 2025. October 7, 2025.

— , “Cipher Mining Commences Bitcoin Mining at Black Pearl Data Center,” Press release June 2025. June 23, 2025.

CoreSite Realty Corporation, “Form 10-K for the Fiscal Year Ended December 31, 2020,” U.S. Securities and Exchange Commission February 2021. Item 1, data center portfolio table, note 2.

CyrusOne Inc., “Form 10-K for the Fiscal Year Ended December 31, 2019,” U.S. Securities and Exchange Commission 2020. Data center portfolio table, note (b).

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.

Deryugina, Tatyana, Alexander MacKay, and Julian Reif, “The long-run dynamics of electricity demand: Evidence from municipal aggregation,” *American Economic Journal: Applied Economics*, 2020, 12 (1), 86–114.

EPRI, “Powering Intelligence 2026: Updated Scenarios of U.S. Data Center Electricity Use and Power Strategies,” White Paper, Electric Power Research Institute, Palo Alto, CA February 2026. p. 14.

Fabra, Natalia, David Rapson, Mar Reguant, and Jingyuan Wang, “Estimating the elasticity to real-time pricing: evidence from the Spanish electricity market,” in “AEA Papers and Proceedings,” Vol. 111 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2021, pp. 425–429.

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.

Gonzales, Luis E, Koichiro Ito, and Mar Reguant, “The investment effects of market integration: Evidence from renewable energy expansion in Chile,” *Econometrica*, 2023, 91 (5), 1659–1693.

Gowrisankaran, Gautam, Ashley Langer, and Mar Reguant, “Energy transitions in regulated markets,” Technical Report, National Bureau of Economic Research 2024.

—, —, and —, “Energy transitions in regulated markets,” *American Economic Review*, 2026, 116 (8), 2928–2961.

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.

Hausman, Jerry, “Mismeasured variables in econometric analysis: problems from the right and problems from the left,” *Journal of Economic perspectives*, 2001, 15 (4), 57–67.

He, Guannan, Soumya Kar, Javad Mohammadi, Panayiotis Moutis, and Jay F Whitacre, “Power system dispatch with marginal degradation cost of battery storage,” *IEEE Transactions on Power Systems*, 2020, 36 (4), 3552–3562.

Ito, Koichiro, “Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing,” *American Economic Review*, 2014, 104 (2), 537–563.

JLL, “North America Data Center Report: Midyear 2025,” Technical Report, JLL August 2025. pp. 2, 5.

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.

Kudela, Peter, Tomas Havranek, Zuzana Irsova, Anna Kudelova, and Vojtech Sikl, “Electricity demand has not become more price-responsive despite ninety years of technological change,” *arXiv preprint arXiv:2607.21285*, 2026.

Lawrence Berkeley National Laboratory and GridTracker, “Queued Up Data File: Interconnection Queue Data through 2025,” Dataset May 2026. CC BY 4.0.

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.

Mansur, Erin T. and Matthew W. White, “Market Organization and Efficiency in Electricity Markets,” 2012. Working paper, Dartmouth College.

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.

NVIDIA, “NVIDIA Ethernet Networking Accelerates World’s Largest AI Supercomputer, Built by xAI,” NVIDIA Newsroom press release October 2024. October 28, 2024.

Rennert, Kevin, Frank Errickson, Brian C Prest, Lisa Rennels, Richard G Newell, William Pizer, Cora Kingdon, Jordan Wingenroth, Roger Cooke, Bryan Parthum et al., “Comprehensive evidence implies a higher social cost of CO₂,” *Nature*, 2022, 610 (7933), 687–692.

Riot Platforms, Inc., “Form 8-K, Exhibit 99.1 (Corsicana Facility update),” U.S. Securities and Exchange Commission April 2024. April 18, 2024.

—, “Riot Announces June 2024 Production and Operations Updates,” Press release July 2024. July 3, 2024.

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.

Samuelson, Paul A., “Spatial Price Equilibrium and Linear Programming,” *American Economic Review*, 1952, 42 (3), 283–303.

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.

TeraWulf Inc., “TeraWulf Announces Full Deployment of 50 MW at the Nautilus Bitcoin Mining Facility Ahead of Schedule,” Press release April 2023. April 20, 2023.

U.S. Energy Information Administration, “About the EIA-930 Data: Hourly Electric Grid Monitor,” <https://www.eia.gov/electricity/930-content/about.html> 2026. Form EIA-930, Hourly and Daily Balancing Authority Operations Report.

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.

Zivin, Joshua S Graff, Matthew J Kotchen, and Erin T Mansur, “Spatial and temporal heterogeneity of marginal emissions: Implications for electric cars and other electricity-shifting policies,” *Journal of Economic Behavior & Organization*, 2014, 107, 248–268.

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). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

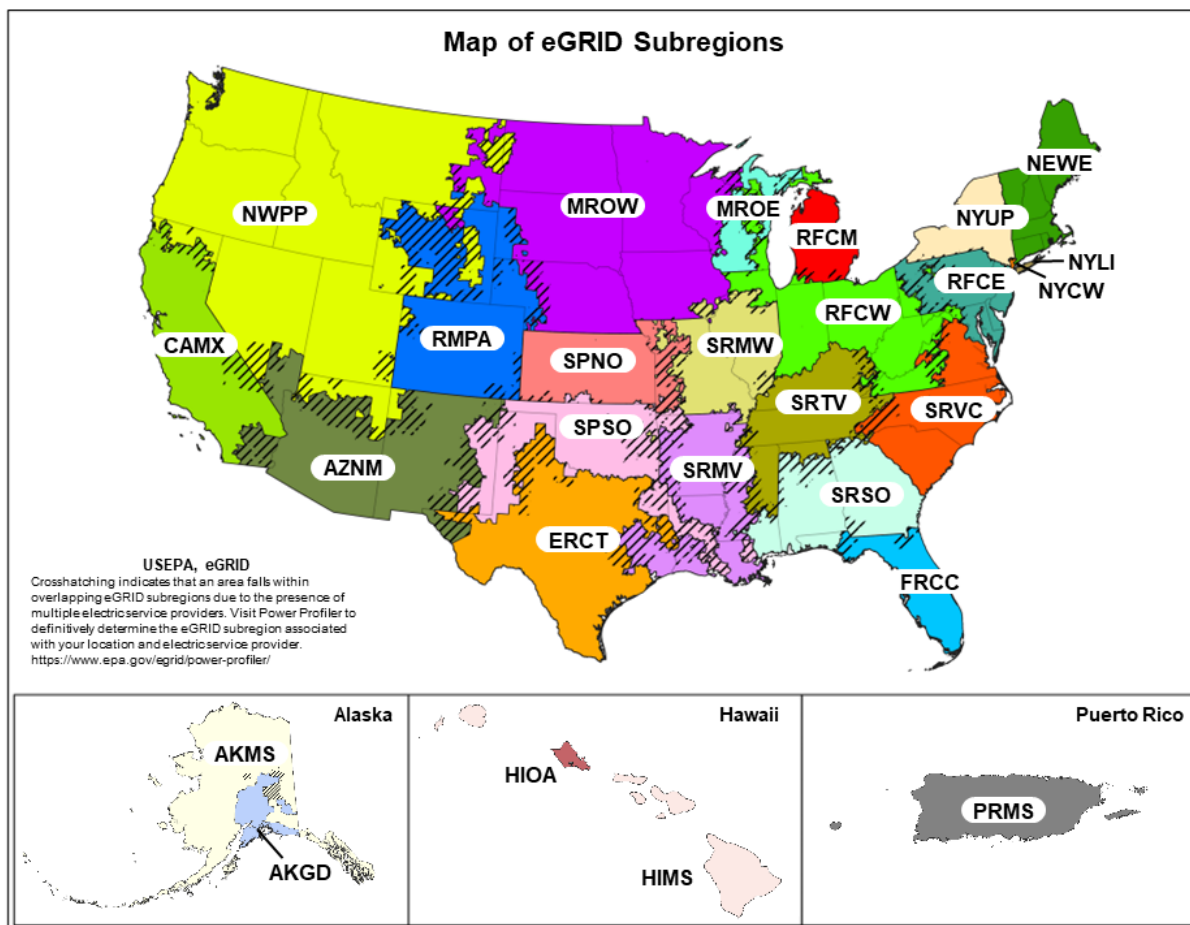
Appendix Table A2: Data Center Effects With and Without Inter-Regional Trade, 2021–2025

	National price effect			Cost	CO ₂	Unserved
	2025 (\$/MWh)	2025 (%)	2021–25 (%)	2021–25 (\$B)	2021–25 (Mt)	energy (TWh)
<i>Mid-Range (Peak)</i>						
Endogenous trade (baseline)	2.66	6.9	5.9	43.7	784.0	0.01
Trade fixed at observed	2.99	7.8	5.7	42.6	787.8	0.14

Notes: Each row re-solves the ex-post dispatch for the observed system and for the scenario's no-data-center counterfactual. *Endogenous trade* is the baseline model. *No trade* sets every inter-regional transfer capability to zero, so each subregion serves its own residual demand with its own thermal fleet and storage. *Trade fixed at observed* follows [Hausman \(2025\)](#) and [Ham et al. \(2025\)](#): each subregion's residual demand is its observed thermal net generation, which holds inter-regional flows at their observed level; the data-center load is removed from it for the counterfactual, and storage is not re-dispatched. Price effects are the difference between the with- and without-data-center wholesale price, weighted by no-data-center demand, in 2024 dollars; annual values average the monthly effects. Cost is the change in thermal variable cost (fuel, variable O&M, permits); CO₂ in million short tons. Unserved energy is load that cannot be met after the mop-up step, in the scenario shown, summed over 2021–2025.

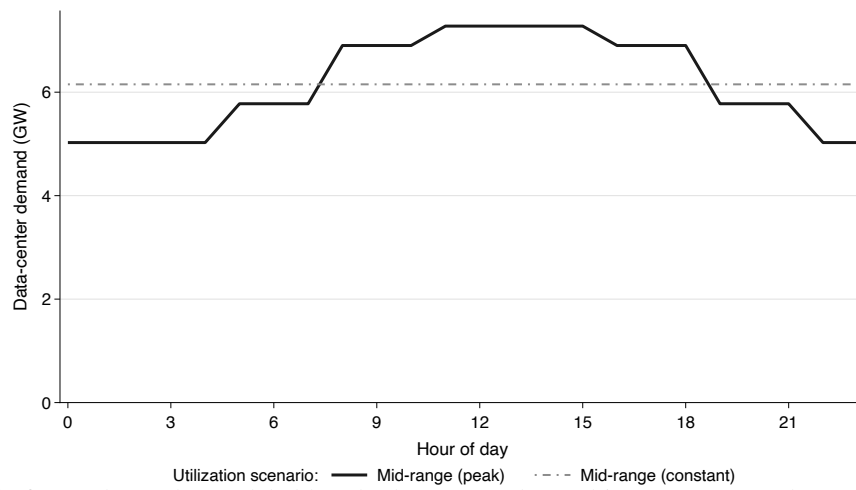
B Appendix: Figures

Appendix Figure A1



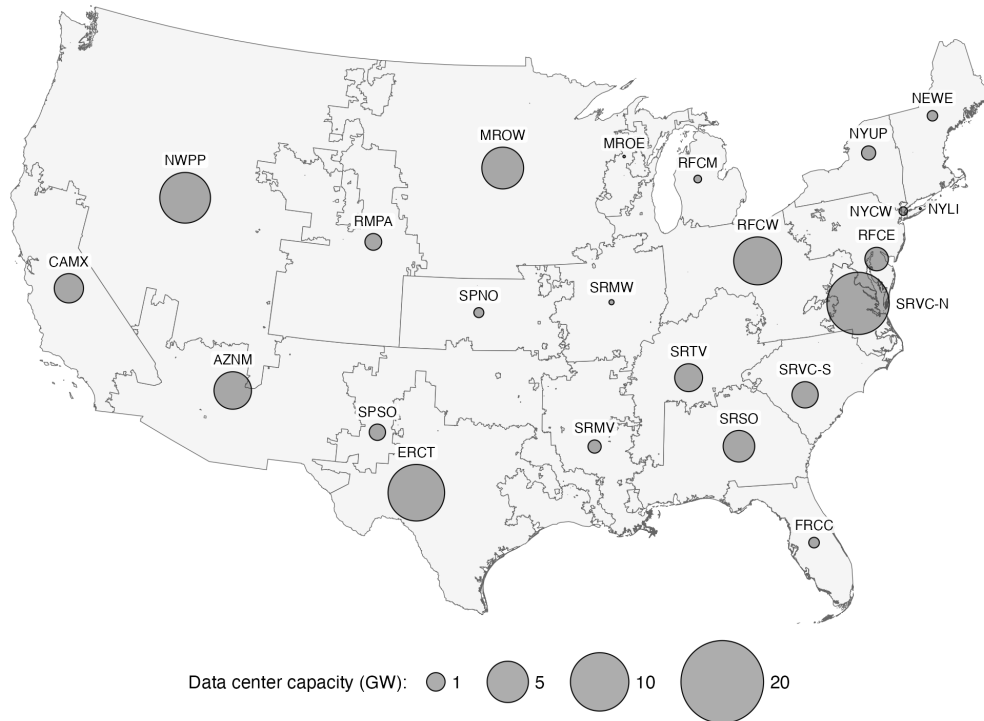
Notes: This map depicts the EPA's eGRID subregions. Our model covers the 22 subregions of the contiguous United States shown here, with SRVC further divided into SRVC-N and SRVC-S, for 23 regions in total Source: Environmental Protection Agency, eGRID.

Appendix Figure A2: Hourly Data Center Demand on December 31, 2025 in SRVC-N under Mid-Range Scenarios



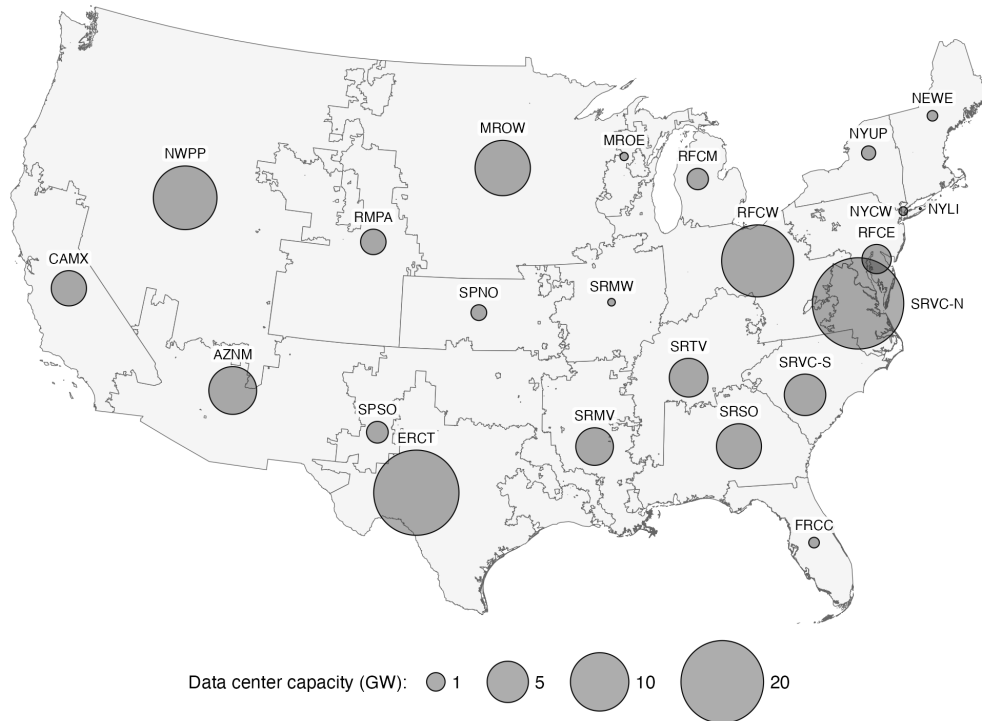
Notes: This figure shows the modeled hourly data center demand in SRVC-N (northern Virginia) under the two Mid-Range scenarios, in GW. Our Mid-Range scenarios are constructed to sum to the same daily energy values while varying in their hourly consumption. The Peak load shape is constructed by data center type, following [EPRI \(2026\)](#).

Appendix Figure A3: Installed Data Center Capacity: 2025



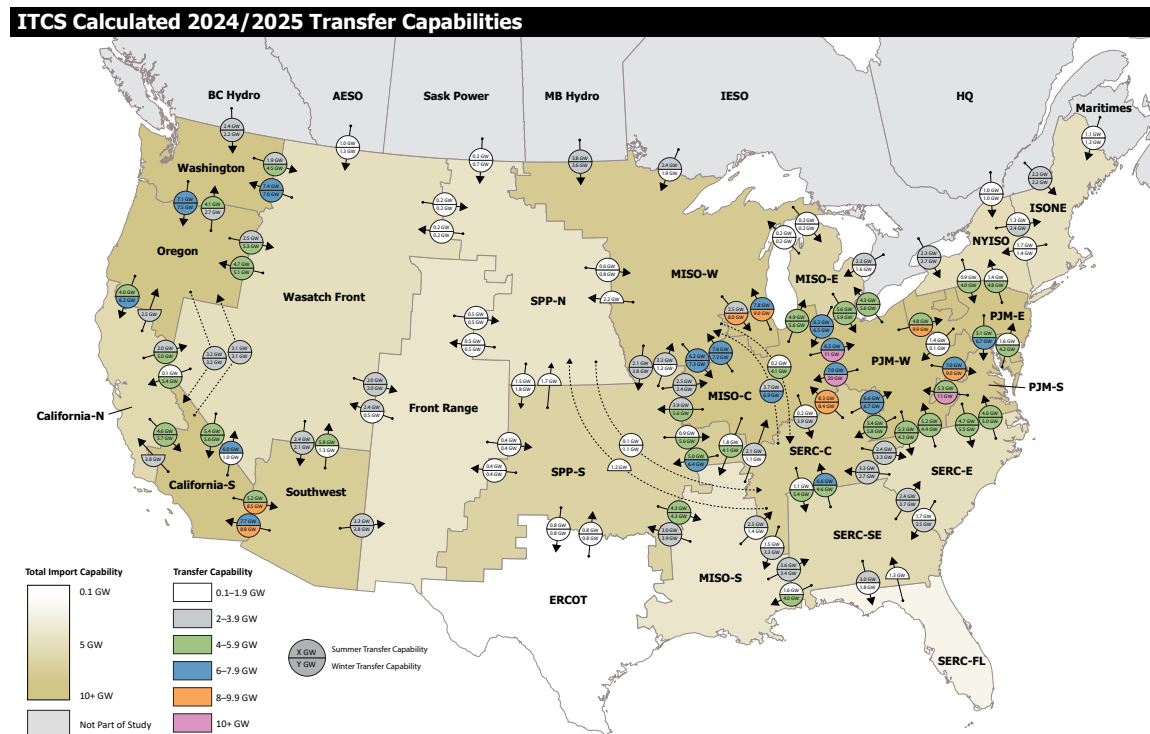
Notes: This map plots the operational data center power capacity as of the end of 2025 by eGRID subregion. SRVC is further split into SRVC-N and SRVC-S. Data are sourced from S&P 451 Research Data Center KnowledgeBase.

Appendix Figure A4: Installed and Planned Data Center Capacity: 2028



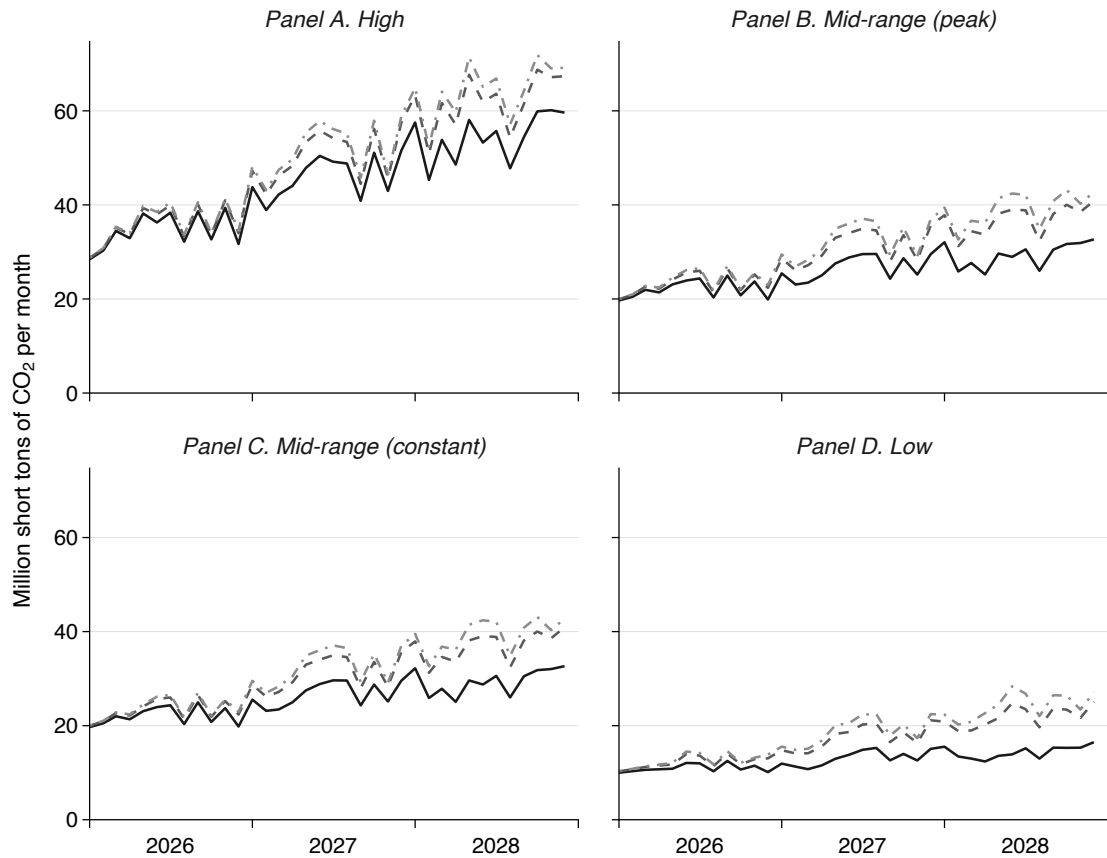
Notes: This map plots the operational data center power capacity as of the end of 2028 by eGRID subregion. SRVC is further split into SRVC-N and SRVC-S. Data are sourced from S&P 451 Research Data Center KnowledgeBase.

Appendix Figure A5



Notes: This map depicts the transfer capability between NERC planning regions for summer and winter of 2025/2025. We map these regions onto eGRID subregions to construct the interfaces in Figure 1. Where the mapping is not one-to-one we supplement this construction with the 99th percentile of observed EIA-930 interchange between balancing authorities. Source: NERC, 2024 Interregional Transfer Capability Study.

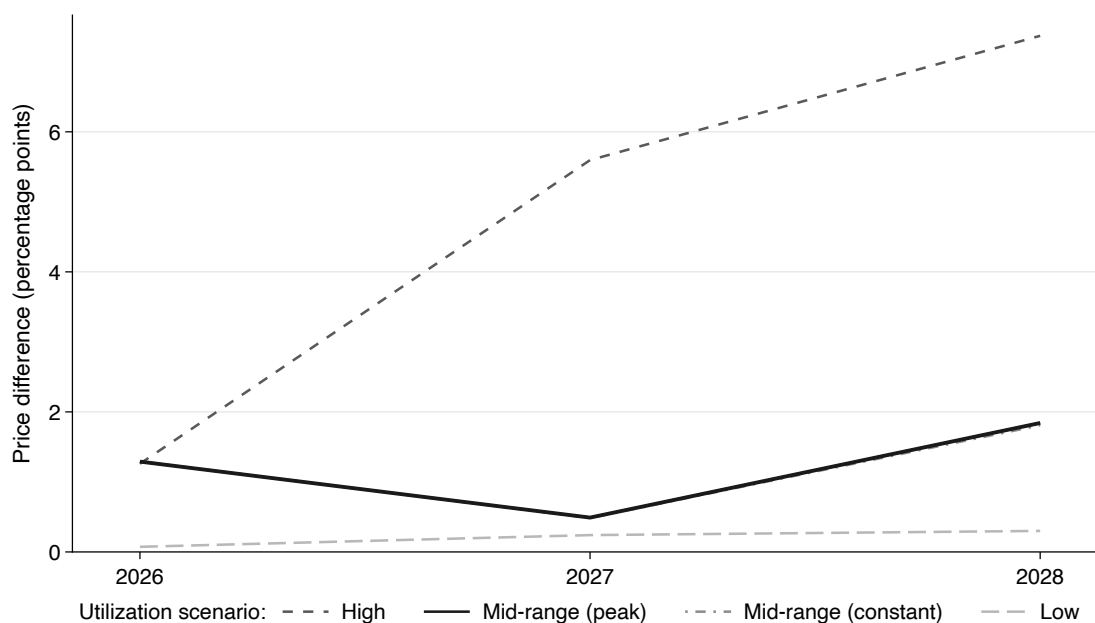
Appendix Figure A6: Forecasted Incremental Carbon Emissions under Alternative Renewables Build-Out, 2026–2028



Renewable build-out: — Full build-out - - - 50% solar, no new wind · - · 25% solar, no new wind

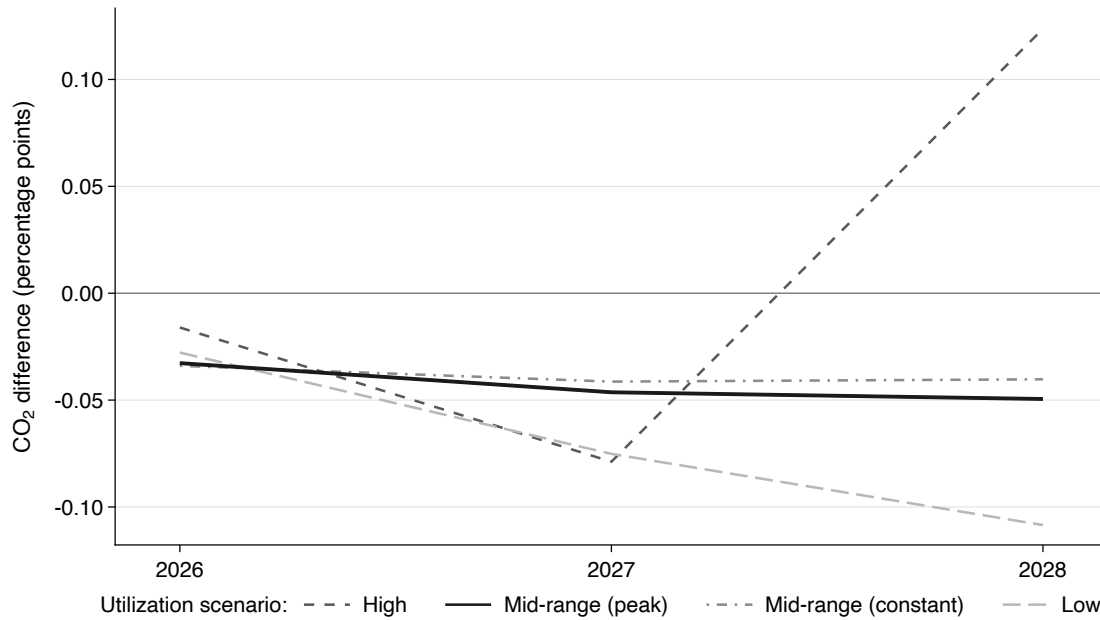
Notes: This figure shows the incremental CO₂ emissions attributable to data centers under our three counterfactual renewable build-out paths, in millions of short tons per month. Each scenario is measured against a no-data-center counterfactual built on the same renewable path.

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



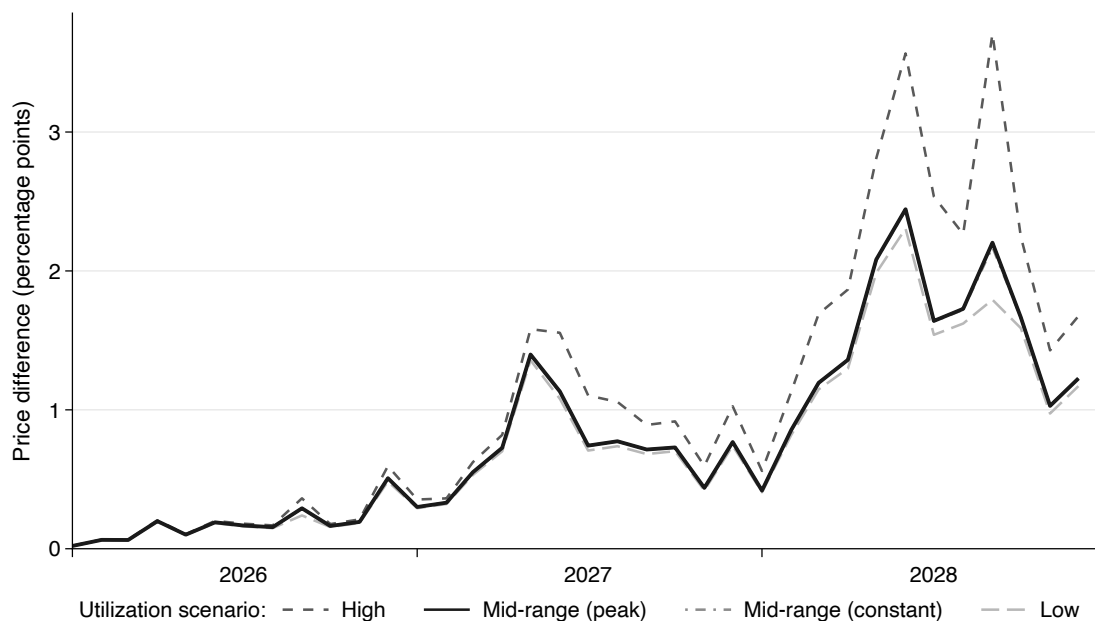
Notes: This figure compares the load-weighted average annual price impacts attributable to data centers relative to our baseline model under the assumption that no planned capacity retirements occur. Through 2028, the vast majority of planned retirements are for coal. For each scenario, we compute the difference, in percentage points, between the price impact where planned retirements remain online in both the scenario and its counterfactual and our baseline price impacts. We aggregate these results to annual levels for clarity, as the individual timing of retirements makes these results less-interpretable at the monthly level.

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



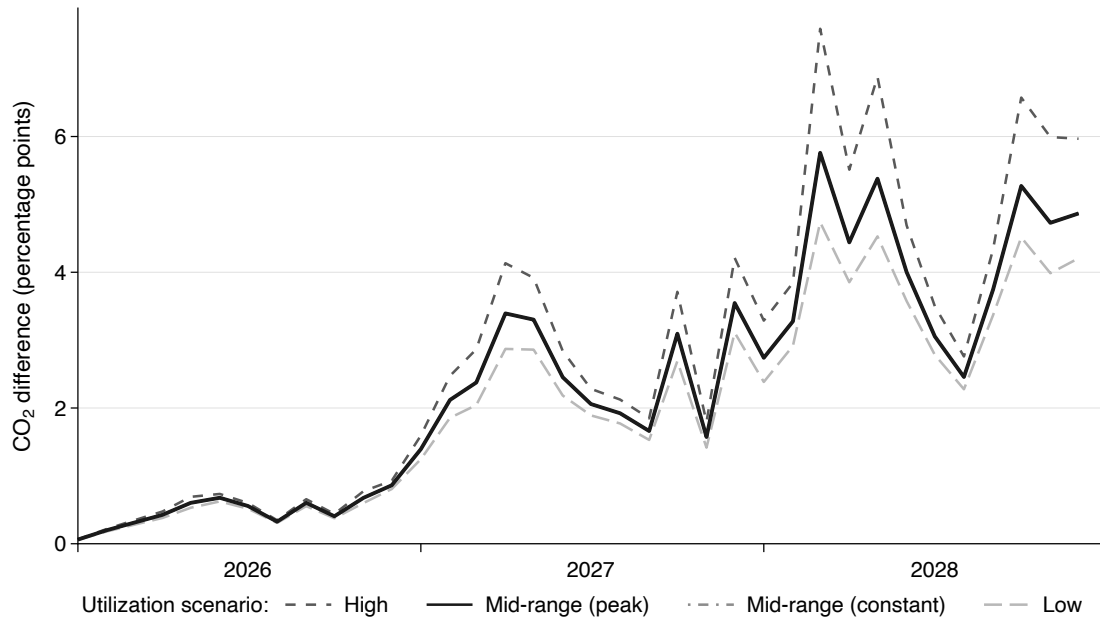
Notes: This figure compares the annual incremental CO₂ impacts attributable to data centers relative to our baseline model under the assumption that no planned capacity retirements occur. Through 2028, the vast majority of planned retirements are for coal. For each scenario, we compute the difference, in percentage points, between the emissions impact where planned retirements remain online in both the scenario and its counterfactual and our baseline price impacts. We aggregate these results to annual levels for clarity, as the individual timing of retirements makes these results less interpretable at the monthly level.

Appendix Figure A9: Difference in Forecast Wholesale Market Price under Endogenous Build-out Assumption, 2026–2028



Notes: This figure compares the load-weighted average monthly data center price impacts relative to our baseline model under the assumption that 25% of planned generation capacity is an endogenous response to load growth caused by data centers. For each scenario, we compute the difference, in percentage points, between the price impact where the no-data-center counterfactual contains only 75% of planned generation capacity and our baseline model, which assumes capacity additions are exogenous to data center build out in the short run.

Appendix Figure A10: Difference in Forecast Carbon Emissions under
Endogenous Build-out Assumption, 2026–2028



Utilization scenario: - - - High — Mid-range (peak) ···· Mid-range (constant) — Low

Notes: This figure compares the monthly incremental CO₂ emissions attributable to data centers relative to our baseline model under the assumption that 25% of planned generation capacity is an endogenous response to load growth caused by data centers. For each scenario, we compute the difference, in percentage points, between the emissions impact where the no-data-center counterfactual contains only 75% of planned generation capacity and our baseline model, which assumes capacity additions are exogenous to data center build out in the short run.

C Appendix: Data Construction

C.1 Thermal Fleet and Fuel Costs

Table A3 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) across technology types (combined cycle versus simple cycle). There is also a significant amount of oil-fired generation capacity, but these units tend to have high marginal costs and are among the last generators to be dispatched.

Appendix Table A3: Thermal Generator Characteristics Statistics

Fuel	N Units	Avg Capacity	Min Capacity	Max Capacity	Avg MC	Min MC	Max MC
Coal	368	432.7	12.6	1294.7	32.4	14.5	85.9
Coke	5	148.3	19.0	277.5	39.2	36.1	48.2
NG	2772	147.1	0.6	853.8	54.9	17.1	506.0
Oil	237	45.4	0.2	548.4	431.9	135.8	861.1

Notes: This table summarizes the capacity and marginal costs of the 3382 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.

Winter delivered gas prices. In several regions, delivered prices rise far above Henry Hub in winter, when pipeline capacity tends to be tighter, and our state-by-month markup fails to capture important price variation. We therefore measure the ratio of Form 923 fuel receipts to Henry Hub prices by eGRID subregion and season. Specifically, we divide the volume-weighted delivered cost by the average Henry Hub price over that month and take the median across our ex post sample. When the winter ratio is at least 10 percent higher than the average of the summer and shoulder ratios, the unit's winter gas price becomes the Henry Hub price times its seasonal ratio. This is particularly important in New England and New York.

C.2 Ex Post Residual Demand

Solar, wind, and hydro. We map the universe of solar, wind, and hydro units to eGRID subregions and collect unit-level nameplate capacities, primarily from the eGRID facility dataset, with supplemental data from EIA-860 on plant locations and balancing authorities where gaps exist. Because eGRID subregions and balancing authorities do not have a one-to-one match, we redistribute observed generation according to the installed generation mix of each subregion. First, we aggregate each balancing authority's hourly generation to the fuel-by-month level and allocate these totals to subregions in proportion to their units' reported EIA-923 monthly net generation. For each balancing authority, we compute an hourly capacity factor for each fuel type as the fraction of that month's total generation observed in each hour of our sample. We assign these capacity factors to each resource located within the corresponding balancing authority. This ensures that the sum of the resulting unit-level hourly generation is consistent with the monthly total net generation from the EIA-930.

Other inframarginal supply. Hourly generation profiles for other inframarginal supply types, such as nuclear, geothermal, biomass, and cogeneration plants, are relatively flat compared with the exogenous generation of wind and solar, or with hydro generation, which tends to follow hourly load. We construct this generation as a monthly residual using the fact that, for each subregion,

$$\text{Total Net Generation} = \text{CEMS Thermal} + \text{Solar} + \text{Wind} + \text{Hydro} + \text{Other}.$$

Similar to how we treat renewables, we apportion this generation according to the reported EIA-923 monthly net generation totals. This step is particularly important, as different fuel types that fall into this category (e.g., nuclear and small natural gas units) can have very different capacity factors. We subtract monthly CEMS totals and monthly renewable totals from total net generation and spread that residual uniformly across hours within the month.

In later years of our sample, the EIA-930 began reporting storage. Even within balancing

authorities with large storage penetration, such as CAISO and ERCOT, storage's total monthly net generation is negligible and slightly negative, because batteries must charge whatever they discharge.

There are well-known reporting inconsistencies between EIA's Form 930 and 923, which we handle on a case-by-case manner. The generation totals reported in 930 cover units the balancing authority meters or whose operations it can see. This results in under-counting when comparing to plant-level totals reporting in the EIA-923, or over-reporting in cases where balancing authorities report generation from units they both observe ([U.S. Energy Information Administration 2026](#)). One extreme example in the data comes from Bonneville Power Administration, a federally regulated agency, with many non-federal plants in its area owned and scheduled by neighboring utilities. From 2021 to 2024, Bonneville's demand exceeds generation net of interchange by approximately 35 TWh per year. For this region, we utilize total monthly generation from EIA-923 instead, and remove that generation from utilities where there is overlap.

C.3 Data Center Load Profiles and Ramp-Up

There is little visibility into how data centers are operated in practice. We reflect this uncertainty by constructing four data center demand profiles that allow us to analyze the effect on wholesale markets 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 from hyperscale, cloud, and neo-cloud data centers equal to the relative market share for training data centers to every hour as these have a 100% utilization

rate.⁵³ Next, we apply type-specific utilization rates to the remaining data center capacity. In our Mid-Range-Constant scenario, we assign this power equally across all hours of the day. In our Mid-Range-Peak scenario, each data center’s electricity consumption fluctuates around its type’s utilization rate, in a load-following pattern, which is constructed to mirror the patterns documented in [EPRI \(2026\)](#). In particular, from 11AM–3PM data centers operate at a percentage above their utilization rate, depending on their type.⁵⁴ In the hours from 10PM–4AM data centers operate at an equivalent percentage *below* their utilization rate. In shoulder hours, data centers ramp their demand up and down in increments that are halfway between their peak and overnight demands.

The last source of uncertainty is the rate at which each data center’s power consumption rises after its stated operation date. This can vary substantially across and within data center types as construction timelines vary and campuses can scale multiple buildings over time. The DCKB data allows us to handle within-campus staggered energization, as it contains separate records for buildings with different operating dates.⁵⁵ Our ramping assumptions primarily capture within building differences across data center types. We assume buildings at each campus ramp up to full capacity at the following rates: Crypto Mines ramp in 3 months, Hyperscale, Cloud, Neo-Cloud, Wholesale and Reseller take 9 months, Telco and System Integrators take 12 months, and Retail and Hosting take 24 months. We apply a linear ramp to their operational capacity at the monthly level until operational capacity equals full capacity.

Crypto miners tend to install equipment and ramp their facilities as soon as possible. Recent 50–150 MW mining phases were fully deployed within two to three months of their first operational dates (e.g. [TeraWulf Inc. \(2023\)](#); [Riot Platforms, Inc. \(2024a;b\)](#); [Cipher Mining Inc. \(2025b;a\)](#)). Much of the hyperscale and wholesale capacity tends to be pre-leased before operations begin, leading to relatively fast deployment once operations begin ([JLL 2025](#)). In addition,

⁵³We do not observe which data centers are dedicated to AI training versus those dedicated to inference work. And, these designations are not necessarily fixed. This motivates our decision to split the most likely training types — hyperscale, cloud, and neo-cloud — into training and non-training shares.

⁵⁴Hyperscale data centers operate at 10% above their utilization rate, colocation data centers operate at +30% and all other types operate at +50%.

⁵⁵In our sample, 83% of hyperscale capacity is mapped to campuses with multiple records. The median hyperscale campus spreads its operating dates across two years.

AI clusters have been documented to deploy within weeks of energization ([NVIDIA 2024](#)). Our 24-month ramp for retail and hosting approximately matches typical lease-up windows reported within the industry ([CoreSite Realty Corporation 2021](#); [CyrusOne Inc. 2020](#)).

C.4 Trade Interfaces

To construct eGRID trade interfaces we take several steps, depending on the overlap between ITCS and eGRID regions. First, where eGRID boundaries and ITCS boundaries align, we simply use the Summer 2024 and Winter 2024/2025 capacities for these eGRID interfaces. Shoulder seasons are assigned the simple average between the summer and winter capacities. For cases where an ITCS region is fully subsumed by an eGRID region, we assume unlimited transmission capacity within the eGRID subregion and across ITCS regions.⁵⁶ When subregions share an ITCS boundary, we give them the same interface capacity, but cap their aggregate trade capacity at the ITCS level.

When regions lack a direct mapping to an ITCS planning area, we utilize the EIA's Form 930 data on interchange between balancing authorities. In these rare cases, such as the mismatch between the AZNM eGRID region and the Front Range and Wasatch Front ITCS planning areas, we sum the 99th percentile of observed interchange between the balancing authorities contained within the relevant eGRID subregions to construct our trade interfaces. In cases where two subregions have multiple interfaces, we take the sum of the interface capacities to construct our interfaces. Finally, we cap each eGRID subregion's total import capacity at the stated ITCS import capabilities, summing import capabilities where ITCS regions are subsumed by an eGRID subregion. In total, we construct 42 interfaces between subregions. We display the resulting interfaces and import capacities in Figure 1.

⁵⁶For example, the ITCS report specifies separate regions for Oregon, Washington, and the Wasatch Front; these regions are all subsumed into the Northwest Power Pool (NWPP) eGrid subregion, so trade is unrestricted between these zones in our model.

Splitting SRVC. We split SRVC into two regions, SRVC-N and SRVC-S to better capture the price effects of the heavy concentration of data centers in Northern Virginia, which largely fall in the Dominion Energy service area. On the demand side, we represent SRVC-N with Dominion’s zone of PJM by using the EIA-930 subregion demand series. We set SRVC-S to the remainder to ensure the total energy reconciles. We largely assign generation and data centers by their geographic locations. SRVC-N is the portion of SRVC that overlaps with Virginia and the 18 northeastern counties in North Carolina, which approximate Dominion’s service area, while SRVC-S makes up the remainder. For units such as the 1.6 GW Mount Storm coal plant, which is located in West Virginia, but assigned to SRVC (its location on Dominion’s 500 kV system means most of its energy will flow to SRVC according to the physics of the grid), we assign Dominion affiliated plants, such as Mount Storm, to SRVC-N, and the rest to SRVC-S. The ITCS report already splits these regions, so we split their interfaces and import capacities accordingly.

In practice, these interface and import constraints act as soft caps that only bind under particularly tight conditions because moving power across interfaces can be costly. As electricity travels across the transmission network, some of that energy is lost. In addition, intra-regional transmission constraints may prevent the dispatch of thermal generation in one subregion from meeting the residual demand in another subregion. There are also persistent transaction costs for trading across the boundaries of adjacent RTOs/ISOs (Mansur and White 2012). Failing to incorporate these costs will cause our model to overstate the amount of trade between subregions.⁵⁷ We account for these frictions with a hurdle rate — a per-MWh charge on flows across interfaces.⁵⁸

C.5 Hurdle Rate Estimation

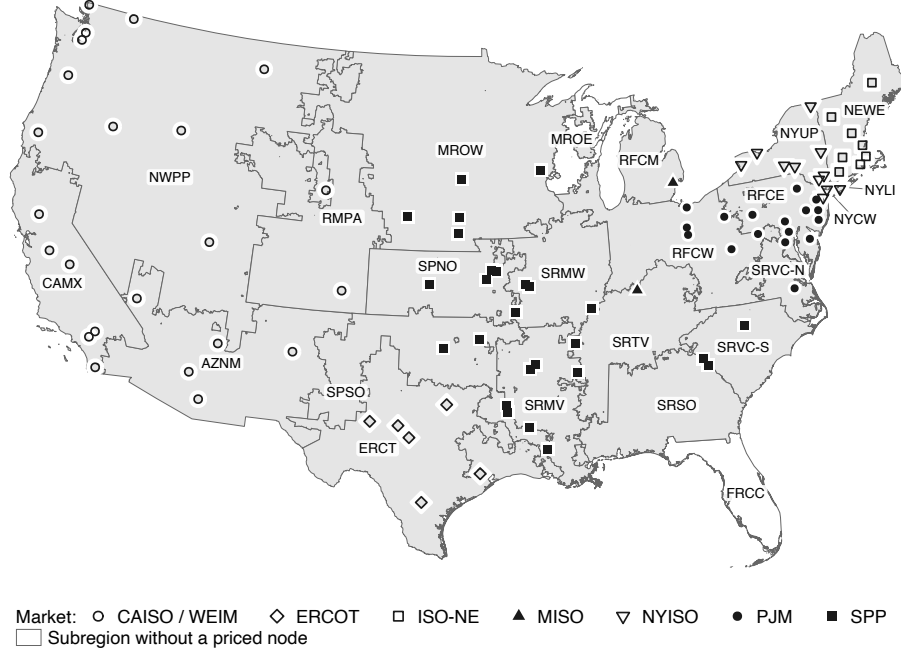
To estimate hurdle rates, we leverage the fact that in a spatial equilibrium with trade over unconstrained links, the price gap per unit between the two areas will reflect the marginal cost of transacting between them (Samuelson 1952). We infer this transaction cost from observed price

⁵⁷Prior research, such as Cicala (2022) and Hausman (2025) show that dispatchable thermal generation primarily responds to nearby demand shocks.

⁵⁸We provide further detail on the hurdle rate calculation in Appendix C.5.

gaps. We collect data on hourly real-time locational marginal prices at 93 distinct pricing nodes in PJM, MISO, NYISO, ISONE, ERCOT, CAISO, and the Western Energy Imbalance Market (WEIM), which spans much of the NWPP region. We map these nodes to geographic coordinates using the S&P Global SNL Energy Dataset. Figure A11 displays the locations of these nodes.

Appendix Figure A11: Map of Price Nodes



Notes: This map displays the 93 pricing nodes used to estimate hurdle rates according to the markets that price them. We display hubs, load zones, and aggregation points at representative locations. Points located in the Western Energy Imbalance Market are shown at each balancing authority's generation aggregation point. There are two regions that do not contain a priced node: FRCC and MROE. Location data comes from the S&P Global SNL Energy Dataset.

We average each node within each subregion according to the dominant ISO/RTO⁵⁹. Let p_{rt}^c and p_{rt}^ℓ denote the congestion and loss components of subregion r 's price in hour t . We represent the hurdle rate on interface e between subregions i and j in season s as

$$w_{es} = \mathbb{E}_s[|p_{it}^c - p_{jt}^c|] + \mathbb{E}_s[|p_{it}^\ell - p_{jt}^\ell|], \quad (16)$$

where $\mathbb{E}_s$ is the average of all hours in season s where subregions i and j are priced. We windorize

⁵⁹SRSO, SRTV, and SRVC-S lie outside of organized wholesale markets. We use the SPP and MISO settlement locations within their boundaries.

these distributions at the 5th and 95th percentiles in order to limit the number of periods where binding constraints between the two periods play a role in their price divergence. We exclude the energy component of all LMPs, keeping the loss and congestion components. Within a market, the energy component will be common and cancel. Across organized markets, this difference will reflect supply and demand conditions, which will be influenced by the shift in data center demand in our model. Therefore, rather than impose this energy gap as an additional cost, we allow our model to arbitrage the difference.⁶⁰ Hurdle rates are estimated as the average absolute difference between two regions since the price gaps between some regions frequently change signs. Hurdle rates are common in both directions across an interface. Two regions in our sample do not contain any pricing nodes: FRCC and MROE. For these regions, we set their hurdles equal to comparably priced interfaces.

We estimate 117 interface seasons in total. Average hurdle rates are \$7.15/MWh with approximately 92% of that composed of the congestion component. Despite only using observed prices to estimate hurdle rates, our trade flows reproduce aggregate observed flows very closely. From 2021 to 2025, simulated hourly net imports across our subregion boundaries are 5.6% of demand while they were reported as 4.7% in the data. As expected, we slightly overstate gross flows over the 11 interfaces that directly map to the EIA 930 data at 1.21 times observed flows.

⁶⁰We find little difference in outcomes if we include this component in the estimated hurdle rates.

D Appendix: Grid-Scale Storage

Along with solar, grid-scale storage capacity is expanding at a rapid pace. From 2021 through 2025 the aggregate U.S. fleet expanded from 4.7 GW of power and 12.5 GWh of energy capacity to 42.8 GW and 118.8 GWh. Our data from Yes and the EIA projects approximately 99.9 GW and 302.9 GWh in 2028. Storage is unique in its role as an electricity arbitrageur.⁶¹ Unlike other resources, storage has neither a fuel cost nor an output schedule dictated by intermittent weather shocks. When a battery chooses to charge and discharge is a dynamic choice, constrained by its rated power (its upper limit on charging and discharging), its energy capacity (the number of hours it can sustain a charge or discharge at its rated power), and its round-trip efficiency. Instead of subtracting observed charge and discharge from each subregion’s residual demand, we represent the battery fleet of each subregions as a single aggregate battery whose charge and discharge are chosen jointly with thermal output and interregional trade in the dispatch program described in Section 4.

D.1 Formulation

Let $x_{rt} \geq 0$ and $y_{rt} \geq 0$ represent battery charge and discharge, in MW, in region r and hour t . Denote the battery’s state of charge at the end of hour t as S_{rt} . The charge and discharge of a battery imposes physical degradation on each battery, κ_r , per MW of charge and discharge. Each region’s battery adds a throughput cost, $\sum_r \sum_t \kappa_r(x_t + y_t)$ to the objective in Equation 2 and enters the region’s power balance constraint in Equation 3. In addition, the battery in each subregion is

⁶¹Hydroelectric 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.

subject to the following constraints:

$$0 \leq x_{rt} \leq P_r, \quad 0 \leq y_{rt} \leq P_r \quad \forall t, \quad (17)$$

$$S_{rt} = S_{r,t-1} + \eta x_{rt} - y_{rt} \quad \forall t, \quad (18)$$

$$0 \leq S_{rt} \leq \alpha_r E_r \quad \forall t, \quad (19)$$

$$\sum_{t \in \mathcal{T}} y_{rt} \leq E_r \quad \forall \text{ days } \mathcal{T}, \quad (20)$$

with rated power (MW) P_r , nameplate energy capacity (MWh) E_r , round-trip efficiency η , and share of available energy capacity $\alpha \in (0, 1]$. Discharging behaves just as our dispatchable thermal generation while charging adds to a subregion's load, which means a subregion's battery is also able to charge from imports and discharge to support exports. Equation 18 tracks the battery's state of charge and de-rates all charged energy by the battery's full round-trip efficiency as it charges. The battery's dynamic marginal cost of energy — the shadow value of carrying an additional MW of charge into the next period — is given by the multiplier on this constraint. Additionally, Equation 19 ensures that the battery's state of charge stays within an operable range, and Equation 20 limits the battery's total discharge to one nameplate cycle per day.⁶²

D.2 Intertemporal Structure and Battery Parameters

Because each region's battery enters the national dispatch problem, we still solve the battery program at the daily level. We allow the state-of-charge to carry into the next day's problem just as we anchor each thermal unit's output at hour 0 each day to its output in hour 23 of the previous day. Each year, the aggregate battery fleet begins half-charged. One advantage to modeling the battery this way is it need not end the day at the state-of-charge it began with. In addition, because we solve the model with a six hour look-ahead period, the battery will carry energy into the next

⁶²In practice, some battery operators choose to cycle multiple times per day. We constrain each regional battery to a single cycle since we do not model uncertainty. We never see the aggregate fleet cycle more than once per-day in the EIA 930 data, so we believe this is a reasonable assumption.

day when prices in those hours justify it.⁶³

We calibrate three parameters to properly model the batteries in our data. First, we set round-trip efficiency at $\eta = 0.85$ since the typical range for grid-scale storage is 0.8 to 0.9. The usable share α_r reflects the reality that battery operators do not cycle through their full nameplate capacity every day.⁶⁴ Utilizing data from EIA Form 930 and ERCOT’s 60-Day SCED Disclosure reports from July 2024 through December 2025, we calculate the fraction of discharge relative to the installed energy capacity in each region. In CAISO, this is 0.68 MWh for each MWh of installed capacity, and in ERCOT is 0.88 MWh given its fleet is made up of shorter duration batteries. We set α_r to these values for CAMX and ERCT and the average of the two in all other regions (given installed battery capacity is much smaller elsewhere). This ensures that fleet cycling each day delivers its typical observed daily energy, given that a battery with perfect foresight tends to over-cycle relative to reality.

Lastly, we include the parameter κ_r , which stands in for physical degradation and variable operating and maintenance costs that enter the battery operator’s marginal cost. These are important frictions to consider, as without them, battery operators would cycle on every small price difference. In reality, these frictions, along with uncertainty, keep them from doing so. Shorter duration batteries tend to cost more per MWh of storage and cycle at higher charge rates. We assume κ_r declines linearly with each region battery’s duration from \$10/MWh to \$5/MWh.⁶⁵ In 2025, κ_r in ERCT is set at \$10/MWh with a duration of 1.58 hours, and \$5/MWh in CAMX, whose fleet averages approximately 3.5 hours. The cost is applied for energy going in and out, which means the throughput cost of each MWh of delivered energy is $\kappa_r(1 + 1/\eta)$ — approximately \$22 in ERCT and \$11 in CAMX.

⁶³We found that look-ahead periods greater than six hours produced very similar results. The six hour look-ahead captures any overnight price spikes that may occur before the predictable morning peak, which makes it unsurprising that using longer look-ahead periods does not change the model’s results.

⁶⁴In reality, they reserve energy for additional ancillary services that batteries excel at providing, and they tend to avoid extremes of the charge and discharge range, as these accelerate the physical degradation of the battery.

⁶⁵Our costs come from [He et al. \(2020\)](#). To remain conservative in our calibration of κ_r , we use values from the lower half of their distribution of estimates.

D.3 Battery Fleet Data

We follow a similar process to construct our ex-post and projected battery panels. For the ex-post period, we utilize Forms 860 and 860m to collect data on battery locations, rated power, and energy capacity. All batteries report a rated power, but approximately 7% are missing their energy capacity. For these, we assume they are 4 hours in CAMX, 2 hours in ERCT, and a uniform random draw between two and four hours in all other regions. For our forward-looking scenarios, we utilize our data from YES Energy along with the data from EIA and follow the same procedure described above. We do not derate battery fleets for outages since we treat each region as having a single battery whose daily energy is calibrated to observed values, which will already contain planned and forced outages.

E Appendix: Ex Post Model Fit

Here, we discuss the goodness of fit of our baseline model by comparing simulated outcomes from our model to observed outcomes from 2021–2025.⁶⁶ We also compare the results from our eGRID subregion model to those from a model aggregated to the regional level and show that, in general, the subregion model appears to match the observed data better.

E.1 Generation Quantities

Table A4 shows that our model using eGRID subregions closely matches first and second moments of observed generation quantities by unit. Additionally, as in reality, the median unit has zero output, since most capacity is offline at any given moment. Simulated generation is more concentrated than observed, as evidenced by the higher standard deviation and lower 75th percentile, which is typical of dispatch models of this kind—units run less often, but closer to full output when they do.

Appendix Table A4: Observed Versus Simulated Generation, Summary Statistics

	Count	First and Second Moments		Percentiles		
		Mean	SD	25th	Median	75th
Observed generation	149,429,424	71.45	136.10	0	0	86.53
Simulated generation	149,429,424	71.26	145.54	0	0	47.62

Table A5 presents the correlations between observed and simulated generation at the unit $\times$ hour, unit $\times$ month, and unit $\times$ year levels. A few mechanisms cause a divergence here between observed and simulated unit-level outcomes. First, we apply uniform derates and stochastic outages across all units, which push simulated capacity from operational capacity in every hour, or make the unit fully unavailable when it was operating in reality. This is further compounded by our use of the previous day’s simulated generation at hour 23 as the initial generation vector at

⁶⁶Given 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.

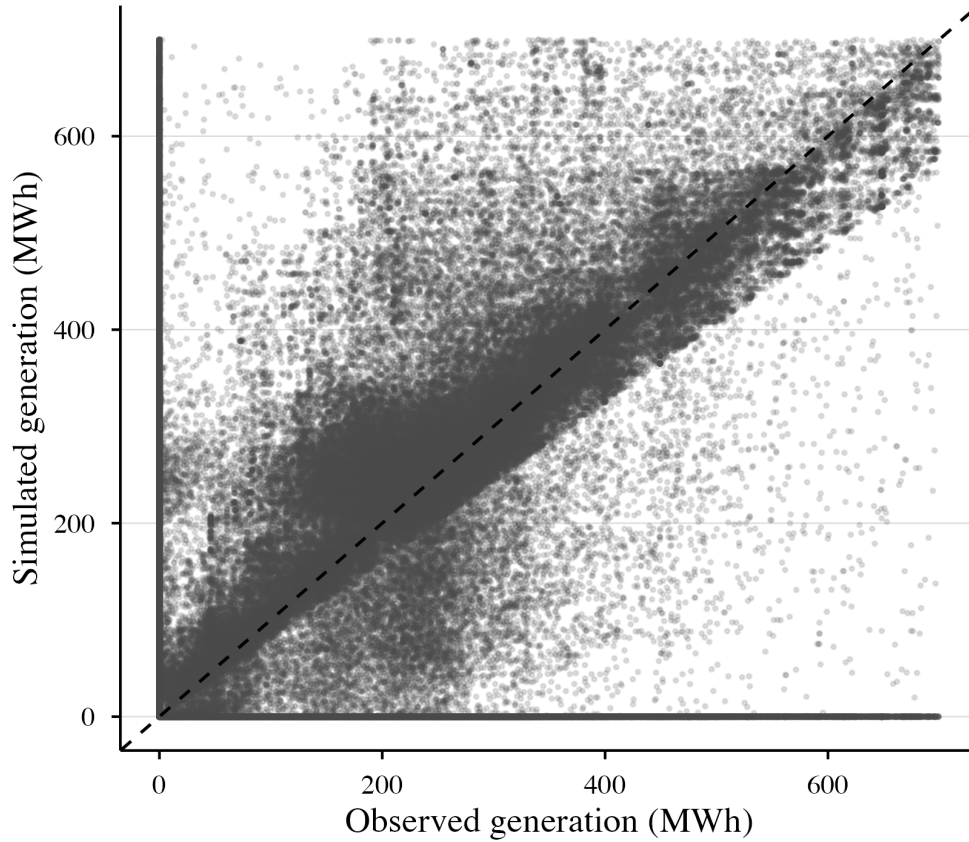
hour zero within a year, which has the potential to create a path dependence in unit-level simulated outcomes that diverges from reality.⁶⁷ Despite this, coal and natural gas units, which consistent almost all units dispatched in our analysis, match observed generation quite well. Figure A12 provides 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 were actually on in reality. Similarly, the points along the y-axis are those observations where the unit was offline in reality, and our model chose to dispatch them. 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 A5.

Appendix Table A5: Correlation Between Observed and Simulated Generation

Fuel	Correlation with Observed Generation		
	Hourly	Monthly (unit $\times$ month)	Yearly (unit $\times$ year)
All	0.634	0.784	0.867
Coal	0.447	0.610	0.737
Coke	0.200	0.298	0.522
NG	0.650	0.811	0.882
Oil	0.075	0.073	0.188

⁶⁷Starting from the observed state would inflate the amount of generation ramping down in the no-data-center counterfactual, leading to an upward bias in our results.

Appendix Figure A12: Model Fit: Hourly Observed Versus Simulated Output



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

E.2 Competitive Wholesale Prices

We also compare our simulated wholesale prices to observed wholesale prices in the seven restructured markets: CAISO, ERCOT, ISO-NE, MISO, NYISO, PJM, and SPP. For each subregion, the observed price is the load-weighted total locational price, which includes its energy, congestion, and loss components at the load zones and hubs that serve the subregion's load. Observed prices also reflect generator bids, within-subregion congestion, and scarcity pricing, which our model does not directly incorporate. This can cause observed and simulated prices to diverge in both hours when there is a glut of renewables and marginal units bid near price floors with dynamic considerations in mind or tight hours when capacity and transmission is constrained. However, in the majority of hours, when there are limited constraints on the system, we should expect prices to

reflect the marginal cost of the generator on the margin.

Table A6 compares simulated versus observed price 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 weighting by the load at each subregion-node contained within wholesale market. Computed over all hours the simulated mean is \$43.93/MWh against an observed \$47.71/MWh (Panel A). Almost all of this gap comes from the tails of the distributions. In Panels B and C, we censor both distributions at their 1st and 99th percentiles. This reduces the pooled average price gap to 0.5%, and the simulated mean in every market (Panel C) falls within \$2.70/MWh of the observed mean. A common issue with merit-order dispatch models is they produce a much flatter distribution. This is further exacerbated by the trade in our model, because the lack of intra-regional constraints leads to an excess of trade relative to observed. However, the inclusion of ramping costs to the model greatly improves the fit in the upper quartile and even in the top decile. Overall, we fit the price distributions well, particularly in the interquartile range, where Figure 6 suggests a large share of the effects lie. In the upper deciles, our estimates will tend to be conservative.

Table A7 reports hourly, monthly, and yearly correlations between simulated and observed prices. After censoring the distributions, the fit across all markets is very, particularly when moving to the monthly and yearly levels. One outlier in the monthly series is CAISO. Western gas prices spiked in December 2022 and January 2023, and this iteration of our model does not capture this. Removing these months, brings CAISO monthly fit improvement in line with the other markets.

F 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. The primary methodological difference between the ex-post and forward-looking residual demand is that we do not observe quantities in our forecast period. We detail the construction below.

Appendix Table A6: Observed Versus Modeled Wholesale Electricity Prices (ISO/RTO level)

	N	Mean	SD	25th pct.	Median	75th pct.	90th pct.
<i>A: All Observations</i>							
Observed price	295,463	47.71	100.25	22.68	33.56	52.67	81.36
Modeled price	295,463	43.93	27.07	28.66	38.65	53.48	73.25
<i>B: Winsorized at 1%/99%</i>							
Observed price	295,463	43.58	36.01	22.68	33.56	52.64	81.19
Modeled price	295,463	43.36	22.81	28.66	38.65	53.48	73.22
<i>C: By ISO/RTO, Winsorized</i>							
CAISO — observed	38,257	53.04	42.67	30.70	45.92	65.46	92.67
CAISO — modeled	38,257	55.71	19.38	41.74	53.69	66.98	81.70
ERCOT — observed	40,423	39.50	40.33	19.43	27.50	44.95	74.85
ERCOT — modeled	40,423	37.86	22.85	23.49	32.82	45.47	68.72
ISO-NE — observed	43,353	57.26	46.03	27.66	41.26	67.77	119.18
ISO-NE — modeled	43,353	59.31	32.84	36.10	50.69	75.04	103.42
MISO — observed	43,358	38.88	23.09	23.39	31.97	46.63	68.95
MISO — modeled	43,358	38.34	17.95	27.73	35.61	46.03	60.45
NYISO — observed	43,356	49.63	36.65	26.68	37.87	57.74	95.36
NYISO — modeled	43,356	51.94	25.74	33.65	45.92	63.52	85.82
PJM — observed	43,358	43.34	28.26	24.52	34.61	52.36	79.29
PJM — modeled	43,358	40.85	18.21	29.88	37.97	48.82	64.22
SPP — observed	43,358	34.20	27.34	18.28	27.13	41.74	68.53
SPP — modeled	43,358	36.68	18.81	25.18	33.62	44.59	59.81

Notes: Observed prices are a load-weighted average of the set of pricing nodes overlapping each eGRID subregion that intersects with an ISO/RTO. Modeled prices are a load-weighted average of the simulated dispatch balance shadow price, in \$/MWh, 2021–2025, (Winter Storm Uri excluded). Panels B and C winsorize each subregion’s distribution, capping values at the 99th percentile and flooring them at the 1st percentile. Censoring affects 2.0% of observed and 1.8% of modeled subregion-hours.

Appendix Table A7: Correlations Between Hourly Observed and Modeled Wholesale Prices
(ISO/RTO level)

	Hourly	Monthly	Annual
<i>A: All Observations</i>			
ISO/RTO Price $\times$ hour	0.408	0.745	0.866
<i>B: Winsorized at 1%/99%</i>			
ISO/RTO Price $\times$ hour	0.678	0.840	0.929
<i>C: By ISO/RTO, Winsorized at 1%/99%</i>			
CAISO	0.642	0.661	0.967
ERCOT	0.597	0.853	0.925
ISO-NE	0.683	0.885	0.953
MISO	0.767	0.964	0.994
NYISO	0.685	0.840	0.897
PJM	0.729	0.948	0.990
SPP	0.647	0.947	0.993

Notes: Pearson correlation of observed and modeled prices. Each ISO/RTO is built from the eGRID subregions overlapping it, weighted by EIA-930 load overlap, with both sides aggregated identically. Monthly/annual first average within ISO $\times$ period, load-weighted. Panels B and C winsorize each subregion's distribution, capping values at the 99th percentile and flooring them at the 1st percentile. Censoring affects 2.0% of observed and 1.8% of modeled subregion-hours, 2021–2025, (Winter Storm Uri excluded).

F.1 Forward Looking Demand Scenarios

To calculate a demand forecast through 2028, we combine data on current hourly production from the EPA’s eGRID, EIA-930, and annual electricity consumption from the EIA Annual Energy Outlook (AEO). Our goal is to construct forward-looking demand that accurately reflects growth in electricity consumption absent projections about data-center-driven growth, maintains realistic inter- and intra-day patterns, and is continuous with observed ex-post demand.

We project hourly subregion demand as the product of a 2025 reference level, a growth trajectory, and an observed hourly shape:

$$\text{Demand}_{r,t} = \text{Ref}_r \times \frac{\text{AEO}_{r,y(t)}}{\text{AEO}_{r,2025}} \times \text{Shape}_{r,b(t)}. \quad (21)$$

Our reference level, Ref_r , is the total demand in subregion r from 2025, net of total data center demand implied by our Mid-Range scenarios, which leverages the Bloomberg/DC Byte utilization rates. The ratio $\frac{\text{AEO}_{r,y(t)}}{\text{AEO}_{r,2025}}$ allows us to calculate year-over-year growth rates from 2021 through 2028, while anchoring them to our reference year. The growth rates use the total forecasted electricity demand by electricity market module (EMM) region taken from the EIA’s 2021 AEO. Crucially, we select the 2021 AEO in order to capture expected demand growth that reflects broader electrification trends, but omits expected growth driven by the AI boom.⁶⁸

To apportion the annual demand number, $\text{Ref}_r \times \frac{\text{AEO}_{r,y(t)}}{\text{AEO}_{r,2025}}$, across hours within a year we utilize our hourly subregion-level demand values from 2021–2025. In each hour, we compute the share of electricity consumed in that subregion relative to the yearly total. 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, $\text{Shape}_{r,b(t)}$, for each calendar month is a random selection from the corresponding donor months. The selected month from the bootstrap draw is

⁶⁸Note that this methodology does not account for data center demand crowding out other sources of electricity demand growth.

consistent across balancing authorities to account for common shocks. Multiplying our annual value by $\text{Shape}_{r,b(t)}$ yields a no-DC hourly load for each subregion for our entire forecast period.

For each of our demand scenarios, we use reported power and operational dates for all operational and proposed data centers from S&P 451 Research, 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 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.⁶⁹ There is also growing local opposition to data centers, leading to the cancellation or delay of several projects.⁷⁰ 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.

F.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.⁷¹

For new builds, we combine proprietary data from YES Energy supplemented with EIA-860m

⁶⁹Analysis by CleanView—a market intelligence firm that tracks data center and other power sector infrastructure development—suggests that approximately 25% of planned capacity is set to connect behind-the-meter: <https://cleanview.co/reports/behind-the-meter-data-centers>.

⁷⁰E.g. <https://www.nytimes.com/2026/06/09/climate/data-center-bans.html>

⁷¹We ignore unit degradation over time. Given our forward-looking scenarios span 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 underestimate costs and emissions across thermals.

data. The YES data has two advantages. First, YES tracks a more comprehensive set of planned generator additions, across all fuel types. Second based on their historic data, YES has computed, for each unit, a probability of its future interconnection by the planned operation date. For all planned units, we derate their capacity by $1 - \text{the probability that unit connects by its planned operating date}$. For the 860m units not in the YES database, we compute a unit-type-by-operating-year probability from the YES data and apply those probabilities to the EIA-860m exclusive units to derate their capacity as well. 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 from the EIA Form 923 that corresponds to each donor month and projected coal prices from the 2025 AEO.⁷²

We construct the inframarginal supply stack in two steps. First, existing solar, wind, and hydro units are assigned their bootstrapped hourly output, using the same draw as the demand profile. Second, we construct the inframarginal supply panel similar to our thermal panel. Using data from YES and EIA-860m, we calculate the probability-derated capacity for all new solar, wind, and hydro units. We then assign them a bootstrapped hourly capacity factor according to their subregion and multiply that by their derated capacity to get their hourly output. We add this

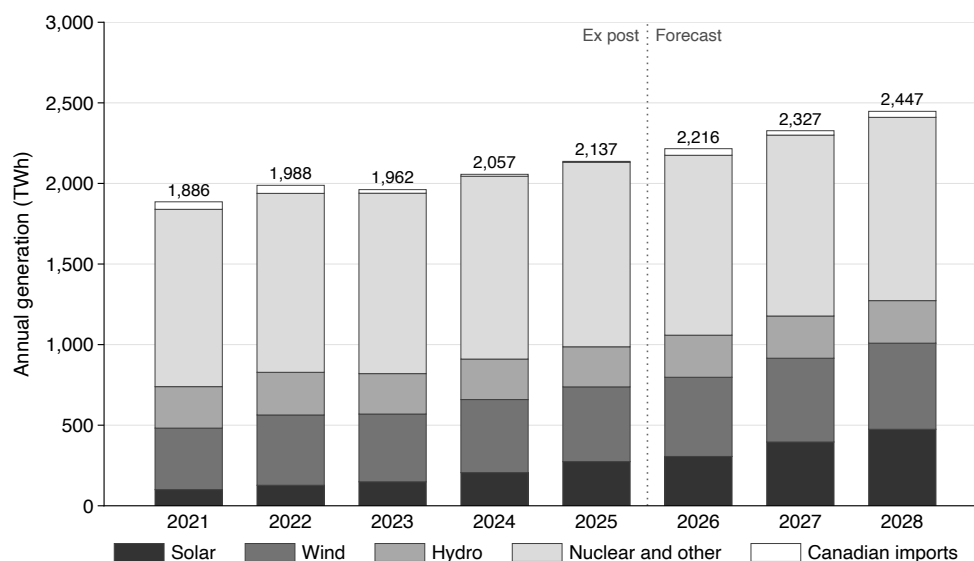
⁷²The AEO does not include projected coke prices for delivered power. Because coal and coke are substitutes, the price of coke is highly correlated with coke. We observe that over our sample, the cost for coke is, on average, \$0.40 per MMBTU higher than coal. Thus, we use the price of coal with a \$0.40 markup for the small number of coke units in our panel.

new renewable supply on top of the existing renewable supply.

Similar to existing renewables, the bootstrapped “other” supply constructed from residual net generation in the ex-post period is taken as given. We handle the expansion of this supply, which primarily consists of some planned nuclear capacity, geothermal, and cogeneration plants, we first compute probability-derated capacities for the set of planned units in our YES and EIA-860m data. Then, for each technology-month, we compute an average capacity factor using EIA-923 data, which reports monthly net generation at the facility-technology level. We multiply each planned unit’s derated nameplate capacity by its uniform hourly capacity factor and add this supply on top of the ex-post “other” supply for each subregion.

Figure A13 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.

Appendix Figure A13: Inframarginal Supply



Notes: This figure displays the annual inframarginal supply across all 23 eGRID subregions, in TWh. This generation is what is subtracted from hourly demand in order to construct the net load our dispatch is solved over. Ex-post years are built from EIA Form 930 observed hourly net generation while forecast years use bootstrapped 930 values plus projected generation from planned capacity additions, which come from EIA Form 860m and YES Energy. The “Nuclear and other” category is primarily made up of nuclear but also includes geothermal, biomass, and cogeneration. Canadian imports are treated as fixed supply.

G 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.⁷³⁷⁴ 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

$$p^{no-DC} = \frac{TFC}{D_1} + AVC(D_1).$$

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

$$p^{DC} = \frac{TFC}{D_1 + \Delta D} + AVC(D_1 + \Delta D).$$

The change in price can therefore be decomposed into the change in average fixed costs and the change in average variable costs as

$$\Delta p = p^{DC} - p^{no-DC} = \underbrace{\frac{TFC}{D_1 + \Delta D} - \frac{TFC}{D_1}}_{\Delta AFC} + \underbrace{AVC(D_1 + \Delta D) - AVC(D_1)}_{\Delta AVC}.$$

⁷³Technically, it maybe more appropriate to consider TFC and AVC as the revenue requirement from the load serving entity coming from fixed costs rather than simply the fixed costs themselves.

⁷⁴Average Variable Costs ($AVC(D)$) can be best approximated by wholesale prices in ISO/RTO regions or the average generation costs in regions without ISOs.

The change in average fixed costs and average variable costs can be written as

$$\Delta AFC = \frac{TFC}{D_1} \left(\frac{-\Delta D/D_1}{1 + \Delta D/D_1} \right) = -AFC_1 \left(\frac{d}{1 + d} \right)$$

$$\Delta AVC = AVC(D_1 + \Delta D) - AVC(D_1) = v \cdot AVC(D_1)$$

where AFC_1 are the average fixed costs (before load growth), $d = \Delta D/D_1$ is the percent increase in electricity demanded from data centers, and $v = \Delta AVC/AVC(D_1)$ is the percent increase in the average variable cost from data centers. The change in averaged fixed costs is negative and represents the fixed-cost spreading effect from new load. If more kWh of demand served by using the “poles and wires” of the system at a higher utilization rate without building new infrastructure, the average fixed cost per kWh will decrease. However, there is an offsetting effect from ΔAVC which represents the fact that the short-run marginal cost curve is upward sloping and in order to meet additional demand it is necessary to use higher marginal cost power plants. The ΔAVC represents the increase in variable costs.

To compare the relative size of these effects, we define the offset ratio as the share of the fixed-cost-spreading benefits from new load that are offset by an increase in variable costs. Formally, this is given as

$$OR = -\frac{\Delta AVC}{\Delta AFC} = \frac{AFC_1}{AVC_1} \frac{1 + d}{d/v} = \frac{1 - s}{s} \frac{1 + d}{\eta}$$

where $s = AFC/(AFC + AVC)$ is the share of the retail rate used to cover fixed costs and $\eta = d/v$ is the price elasticity of electricity supply. Note that the elasticity, η , shows how responsive supply is relative to the relevant energy cost on retail bills, AVC . In wholesale markets this is the wholesale price while in vertically integrated markets the average generation cost is a better measure.

G.1 Calibrating the Fixed Cost Share

The threshold condition in Appendix G compares the share of the retail bill attributable to fixed costs, s , with the break-even share $s^* = (1 + d)/(\eta + 1 + d)$. The elasticity η and the data center load share d come from our dispatch model and are defined for eGRID subregions. The share fixed cost share s must instead be calibrated. To do so, we measure the variable-cost share of retail rates, set s equal to one minus that share, and aggregate the state measures to subregions. We measure the variable-cost share of retail revenue separately for wholesale markets and states without wholesale markets.

States with organized wholesale markets. For the jurisdictions predominately served by an ISO/RTO (CAISO, ERCOT, ISO-NE, MISO, NYISO, PJM, or SPP), we follow [Borenstein and Bushnell \(2022\)](#) and use the wholesale energy price as the short-run marginal cost of producing electricity, and therefore the relevant variable cost measure for retail rates. We match each state to a pricing-hub (or set of pricing hubs) and measure the average annual wholesale price as the load-weighted average. In New England and New York the pricing zones coincide with state boundaries. For states with no pricing point of their own, a nearby location in the same ISO stands in; for example, the MISO Minnesota Hub is used for Iowa, Wisconsin, and the Dakotas. We then find the average retail rate for each state using EIA-861 data and dividing total revenue from sales to ultimate customers by total quantity of sales to ultimate customers. We determine the variable-cost share as the wholesale price divided by the retail rate, and the fixed-cost share as one minus the variable cost share. Finally, we pool estimates across 2021–2024 in order to find a smoother longer-run estimate.⁷⁵

States without organized wholesale markets. For the remaining states, which are mostly vertically integrated, we use data on investor FERC Form 1 filings for 2021–2024 (accessed via the Catalyst Cooperative’s PUDL tool). For each state we pool all Form 1 respondents i and years and

⁷⁵For Texas, we exclude 2021 estimates due to Winter Storm Uri.

compute

$$s_j = 1 - \frac{\sum_{i,t} (\text{Fuel}_{it} + \text{Purchased power}_{it})}{\sum_{i,t} \text{Operating revenue}_{it}}. \quad (22)$$

For a regulated utility, operating revenue approximates the full revenue requirement, including the return on rate base, so this measure corresponds directly to the object in the theoretical model. We use a state only if Form 1 respondents serve at least half of its retail load and at least half of that load is bundled service.

Aggregation to eGRID subregions. Each state is linked to the subregion or subregions in which its load sits. The subregion share is rebuilt from the pooled dollars of its states rather than averaged across state shares:

$$s_r^c = 1 - \frac{\sum_{j \in J_r} \sum_t w_{jt} q_{jt}^c}{\sum_{j \in J_r} \sum_t p_{jt}^c q_{jt}^c}, \quad (23)$$

and the analogous sum of Equation (22) is used for Form 1 subregions. Each subregion turns out to draw entirely on a single source. A state spanning several subregions enters each with its statewide totals. As a result, NYUP, NYCW, and NYLI share New York's value, and Pennsylvania enters both RFCE and RFCW. We hold s fixed at its 2021–2024 value when evaluating both 2025 and 2028. Appendix Table A8 reports the result. A total-sector share is available for 22 of 23 subregions; SRTV is missing because Kentucky and Tennessee are neither in wholesale markets nor do FERC Form 1 filers serve half of the load (the publicly owned TVA does).

Appendix Table A8: Calibrated Fixed-Cost Share by eGRID Subregion

Subregion	States	Source	s (total)	s (industrial)
AZNM	AZ, NM	Form 1	0.629	–
CAMX	CA	Wholesale	0.729	0.644
ERCT	TX	Wholesale	0.440	0.165
FRCC	FL	Form 1	0.692	–
MROE	WI	Wholesale	0.703	0.569
MROW	IA, MN, ND, NE, SD	Wholesale	0.656	0.545
NEWE	CT, MA, ME, NH, RI, VT	Wholesale	0.739	0.631
NWPP	ID, MT, NV, OR, UT	Form 1	0.587	–
NYCW, NYLI, NYUP	NY	Wholesale	0.721	0.325
RFCE	DC, DE, MD, NJ, PA	Wholesale	0.652	0.470
RFCM	MI	Wholesale	0.681	0.471
RFCW	IN, OH, PA, WV	Wholesale	0.599	0.405
RMPA	CO, WY	Form 1	0.614	–
SPNO	KS	Wholesale	0.736	0.626
SPSO	OK	Wholesale	0.506	0.257
SRMV	AR, LA, MS	Wholesale	0.585	0.387
SRMW	IL, MO	Wholesale	0.631	0.495
SRSO	AL, GA	Form 1	0.690	–
SRTV	(KY, TN)	–	–	–
SRVC-N	VA	Wholesale	0.465	0.297
SRVC-S	NC, SC	Form 1	0.684	–

Notes: s is one minus the variable-cost share of retail revenue, pooled over 2021–2024 and over the listed states. Texas pools 2022–2024. Washington (NWPP) and both SRTV states are excluded because they fail the FERC Form 1 coverage screen. The Form 1 measure has no customer-class split.

H Appendix: The Social Planner’s Problem

We compare the realized geography of data center compute relative to the siting that would result, hour-by-hour, if a Social Planner could reallocate data center load across the US in order to minimize short-run operating cost. The Planner solves the same program from Section 4 with one modification: the data center load in each eGRID subregion enters as a choice variable rather than as an exogenous state.

Let d_{rt}^0 denote region r ’s residual demand in hour t in the no-data-center counterfactual, and let $X_t = \sum_r (d_{rt} - d_{rt}^0)$ denote the total national data center load in t under a given utilization scenario. To minimize total system costs, the Planner selects q_{it} , f_{et}^+ , and f_{et}^- as in the original program, and now also each region’s data center allocation $x_{rt} \geq 0$:

$$\min_{q,f,x} \sum_{i \in I} mc_{it} q_{it} + \sum_{e \in E} (w_e^+ f_{et}^+ + w_e^- f_{et}^-) \quad (24)$$

$$\text{s.t.} \quad \sum_{i \in I_r} q_{it} + \Phi_{rt} = d_{rt}^0 + x_{rt} \quad \forall r, t \quad (25)$$

$$\sum_r x_{rt} = X_t, \quad x_{rt} \geq 0 \quad \forall r, t \quad (26)$$

Equation (24) is left unchanged from the baseline model. The power balance constraint in (25) now requires that local generation and net imports must exactly meet the no-data-center residual demand plus any planner-allocated data center load. The constraint in (26) requires that aggregate allocated data center load must equal national load in every hour.⁷⁶ Lastly, the constraints in Equations (4)–(8) from the baseline model hold just as before.

The solution to each hour produces an allocation of data center load that minimizes the national cost of serving compute subject to generator and system constraints. That is, low short-run marginal cost of supply is not sufficient to attract data center compute; the solution must still respect interface and import capacities and generator ramping constraints, which require dynamic feasibility, given the program is solved at the daily-level just as the baseline. The solution to this problem can be seen as an upper bound on the value of geographic flexibility, as it assumes each region has enough data center capacity to serve any allocation of compute.

⁷⁶In each hour, we assume national data center load is fixed. To isolate the benefits of geographic flexibility, we do not allow for temporal flexibility, as this problem is considered elsewhere in the literature (Norris et al. 2025; Knittel et al. 2025; Ross and Ewing 2026)