Electricity System Reliability Costs of Data Centers

Zoey Yiyuan Zhou * Fanyu Wang†

Oct 20, 2026

Abstract

Data centers add large, continuous electricity loads at specific points on the grid. Beyond the energy they consume, serving these loads reliably requires the power system to hold enough capacity, reserves, and voltage support where they connect. We study these reliability costs: how the system adjusts to keep serving data center load reliably, how high the cost is, and how it is allocated. We combine facility-level data on U.S. data centers with data on power plant operations and emissions, generator retirement plans, and wholesale transactions and contracts, and estimate stacked difference-in-differences models around data center activation. We find that data center load draws on the system's reliability margin. It is met largely by peaking units, which normally stand ready for demand peaks and contingencies: after activation, a peaking unit generates 33.9 MWh more per day and emits 23.7 more tons of CO_2, while other fossil units show no significant change. The system also keeps existing capacity available for longer: the annual probability that a fossil generator postpones its planned retirement rises by 0.52 percentage points. Payments to fossil generators rise for both energy and availability: their quarterly energy revenues rise by $3.8 million and their capacity revenues by $359 thousand per generator and zone, and they hold more contracts for capacity and other reliability products. Payments for voltage support also rise, which manifests data centers' uniquely strict requirement for stable voltage. These effects are larger in quarters of extreme weather, when the system is already under stress. The cost of serving data centers therefore includes the cost of maintaining reliability with an older, more polluting fleet, a cost that energy prices alone do not reveal.

JEL Codes: Q41, L94, L51, Q48

*Climate School, Columbia University. Email: yz2851@columbia.edu.
†School of International and Public Affairs, Columbia University. Email: fw2397@columbia.edu.

1 Introduction

Reliable electricity provision requires maintaining a balance between generation and consumption at all times. The system must be able to meet demand during both ordinary operating conditions and periods of stress, such as demand peaks or equipment failures. This requires spare generating and transmission capacity, reserves that can respond to unexpected disruptions, and services that maintain stable frequency and voltage. The expansion of cloud computing and artificial intelligence makes these requirements increasingly consequential. Data centers add large, sustained electricity loads at particular locations and depend on uninterrupted service and stable voltage. By using capacity that would otherwise provide a buffer against disruptions, these loads can increase the resources and expenditures needed to maintain reliability. The cost of accommodating a data center therefore includes both the cost of producing the electricity it consumes and the cost of keeping the system reliable as demand rises. These reliability costs can be concentrated locally: spare generation elsewhere may offer little protection if transmission constraints prevent power from reaching the new load, while voltage support must often be supplied nearby. This paper examines the adjustments needed to serve these loads reliably and how the resulting expenditures reach retail customers.

Although interconnection procedures can slow down data centers' access to the grid, new generation and transmission capacity may still fail to keep pace with the demand from facilities that begin operating (PJM Interconnection, 2026). When supply cannot adjust immediately, additional demand affects the power system through two margins. First, it increases the operating cost of producing electricity. Higher demand moves dispatch up the merit order, increasing the utilization of higher-marginal-cost generators, particularly peaking units. Second, it can create additional expenditures needed to maintain reliability. A system with ample spare generation and transmission capacity may be able to accommodate a new data center using existing resources. The same load addition can have a very different effect when reserve margins are narrow or local transmission constraints bind. In that case, maintaining reliability may require the system operator to procure additional capacity and reliability services, take out-of-market actions, or retain a generator that would otherwise retire. The consequences of data center growth thus depend not only on the size of the new load, but also on the available margin for accommodating it.

The mechanisms through which reliability expenditures arise differ across electricity markets. In a centralized capacity market, additional load increases the quantity of capacity that must be procured. When the supply of capacity is limited, it may also bring a more expensive resource into the auction and raise the clearing price paid to all cleared capacity. Total capacity expenditure can therefore rise through both the quantity procured and the price applied to that capacity. Where reliability depends on a particular local generator, additional load may instead cause the system operator to postpone its retirement through a cost-based retention agreement. Higher load can also increase the use of operating reserves, voltage support, uplift, and other services required to operate the grid securely. These responses are likely to be more consequential when weather-driven demand, generator outages, or transmission constraints place additional stress on a system with limited spare capacity.

How these expenditures are allocated requires separate attention. Data centers pay for the electricity they purchase and may bear some of the costs of connecting to the grid. Other expenditures, however, support the reliability of a broader system. New generation capacity, transmission upgrades, ancillary services, and the continued availability of existing generators may benefit other users as well as the new load that precipitated them. Consistent with these shared benefits, many reliability expenditures are recovered across load in a transmission zone rather than assigned exclusively to an individual customer. Higher fuel and purchased-power costs may similarly pass through utility adjustment clauses to retail customers. Broad cost recovery does not by itself imply inefficient cost shifting. It does, however, mean that the incidence of data center growth depends on the market and regulatory rules through which system costs are recovered.

We examine this propagation through three margins. First, we study how data center entry changes the operation of the existing generation fleet, including the use of peaking units and the associated emissions. Second, we examine whether concentrated load growth changes retirement decisions and increases expenditures on capacity, ancillary services, and agreements that retain generators for reliability. Third, we trace whether higher wholesale procurement and reliability costs enter the adjustment mechanisms through which utilities recover fuel and purchased-power expenses from retail customers. This framework connects the physical response of the grid to reliability procurement and, ultimately, to retail cost recovery. Because reliability services and grid investments may generate benefits beyond the new load that precipitates them, we interpret broad cost recovery as an incidence result rather than, by itself, evidence of inefficient cost shifting.

Our results show that data center entry increases the use of existing peaking units. After activation, a peaking unit in a treated zone generates 33.9 MWh more electricity per day, burns 398 MMBtu more fuel, and operates for 0.31 additional hours. We find no significant change among non-peaking units. CO_2 emissions from peaking units rise by 23.7 tons per unit-day, while emissions per MWh remain unchanged. Data center entry therefore raises emissions by increasing peaker output rather than the emissions intensity of generation.

Data center entry also delays generator retirements. The annual probability that a planned retirement is postponed rises by 0.28 percentage points for all generators and by 0.52 percentage points for fossil generators, with increases among both peakers and non-peaker fossil units. A broader measure that also includes withdrawn retirement plans rises by 0.80 percentage points for fossil generators and by 1.19 percentage points for non-peaker fossil units. Completed retirements do not change significantly during the sample period; the response instead appears in forward-looking retirement plans, which keep existing capacity available for longer.

These changes are reflected in wholesale transactions. After activation, fossil sellers receive an additional \$3.774 million in energy-sale revenues and sell 122,939 MWh more electricity per seller-zone record and quarter. Their capacity-sale revenues rise by \$359 thousand, and they hold more capacity and reliability contracts. Evidence for other reliability services is less conclusive. The reactive-power event studies contain one significant pre-activation coefficient, so we interpret the subsequent increase in reactive-power payments as suggestive. The annual weather interactions show no differential pre-trends, although several quar-

terly paths diverge before activation. In the annual specification, a one-standard-deviation heating-and-cooling degree-day anomaly adds 47,034 MWh to fossil sellers' post-activation energy sales, $8.3 thousand to reactive-power payments, and $16.3 thousand to ancillary-service payments per record and quarter. Together, the results show that data center growth affects not only electricity consumption, but also generator utilization, retirement decisions, and the payments required to keep capacity and reliability services available.

Our paper relates to a growing literature on how data centers affect what electricity consumers pay. Meeks et al. (2026) and Scalera et al. (2026) show that data center entry raises utility retail prices. Meeks et al. (2026) trace the increase to infrastructure investment, larger authorized rate bases and more frequent rate case filings, with larger effects where reserve margins are low and fossil generation is on the margin; Scalera et al. (2026) attribute it mainly to higher peak demand and find that households bear a disproportionate share. Mamkhezri et al. (2026a) and Mamkhezri et al. (2026b) show that the effect of entry on local marginal prices depends on where the load sits relative to transmission constraints. These papers study the end points of the price chain, local marginal prices and retial prices. We study the steps in between, focusing on the cost of maintaining grid reliability, which local marginal price excludes while retail prices bundle with other charges. We also measure how fuel and purchased-power adjustment clauses pass costs to customers, a channel that is updated outside rate cases that the existing literature has investigated.

Our paper also complements work on how electricity supply changes to meet the data center demand. Darmouni et al. (2026) project that the planned build-out will be met largely by gas plants, raising emissions, while Heiss et al. (2026) find that expected data center demand mainly spurs investment in renewables and storage; Du and Guo (2026) show that limited grid capacity can crowd out data center entry. These papers project future outcomes or study entry, of new plants or new data centers. We instead measure how the existing fleet respond to the increased load: whether incumbent fossil plants run more and emit more, and whether retirement schedules change.

More broadly, we build on the literature on how electricity markets and regulation determine the cost of serving demand. The cost and emissions of additional demand depend on which plants are on the margin (Holland and Mansur, 2008; Graff Zivin et al., 2014; Callaway et al., 2018), and losing a large plant raises costs through more expensive replacement generation and transmission constraints (Davis and Hausman, 2016); we study the mirror image, a large concentrated increase in demand. Reliability is not fully paid for by energy prices, which motivates capacity and other out-of-market payments (Joskow and Tirole, 2007; Reguant, 2014), and regulation determines how costs are incurred and passed through to customers (Fabrizio et al., 2007; Cicala, 2015; Fabra and Reguant, 2014). We connect these strands by measuring how a demand shock propagates through reliability payments and regulated cost-recovery mechanisms.

The rest of the paper is organized as follows. Section 2 explains how the grid serves new load and maintains reliability, and describes our data. Section 3 presents the empirical strategy. Section 4 reports the results, following the order of the adjustments described above, and Section 6 concludes.

2 Background and Data

2.1 Data centers and the electricity system

Data centers draw large amounts of power, run at high utilization around the clock, and connect to the grid in large blocks rather than as gradual growth spread across many customers. Figure 1 shows the data centers in our sample, which includes all operating facilities except cryptocurrency mining campuses. Operating capacity grew from about 10 GW in 2014 to about 46 GW by the end of 2025, and more than half of the capacity added since 2014 came online in 2023 or later. Although a few states account for much of this growth, the build-out is spread across the country (Figure 2). Virginia, Oregon, Texas and Ohio account for about half of the capacity added since 2014, but 41 of the 48 contiguous states added some capacity, and 29 added at least 100 MW.

2.2 How new load is served

Demand is met by dispatching generators in order of their marginal cost given the constraints on factors like transmissions, ramping rate, minimum up or down time, etc, in both organized markets and vertically integrated utilities. Additional load moves the system up this merit order, so that peaking units, typically gas combustion turbines and older steam units that usually run only in high-demand hours, are dispatched more often. Because many of these units are less efficient, fuel use and emissions rise more than output. In addition, facing the sustained load growth, the electricity system may need to keep plants that would otherwise retire to meet their resource-adequacy obligations.

2.3 Reliability mechanisms

Keeping the grid reliable requires more than producing enough energy each hour, and most of the payments for it are made outside the energy price. *Capacity* payments compensate generators for being available at the system stress conditions. *Operating reserves* are capacity held unloaded to respond within several minutes to around half an hour if a generator or line fails. *Regulation* balances supply and demand second by second, aiming to maintain a safe frequency level. *Uplift*, or make-whole, payments cover the cost of units the operator commits for reliability when market prices do not. Not every market has every mechanism, and those that do may pay for it through a market, a cost-based tariff rate, or not at all (Table IA1; Appendix A).

Reactive supply and voltage control deserves particular attention. Reactive power holds voltage at the level equipment is designed for, and because it cannot travel far, it must be supplied close to the load. Data centers are unusually sensitive to voltage: their power electronics switch to backup power when voltage dips even briefly, whereas most residential and industrial loads ride through. A cluster of data centers therefore requires tighter local

voltage control and raises the risk that hundreds of megawatts of load drop at once (North American Electric Reliability Corporation, 2026).

A further reliability mechanism compensates generators for postponing retirement. When a proposed retirement would leave the system unable to meet reliability standards, the system operator can arrange for the unit to remain available until transmission upgrades or replacement resources resolve the problem. The owner receives compensation under an approved agreement, with costs recovered from designated customers according to the applicable allocation rules. Terminology varies across markets: examples include reliability-must-run (RMR) agreements, System Support Resource (SSR) agreements in MISO, and Reliability Support Services Agreements (RSSA) in New York. We use "reliability retention agreements" as an umbrella term. These agreements connect retirement decisions to cost allocation: they preserve existing capacity while creating charges that may be shared across customers beyond the new load.

2.4 From system costs to retail rates

In states with rate-regulated utilities, retail rates have two layers. Base rates recover capital investment and fixed operating costs and are reset in periodic rate cases. A separate power supply cost recovery charge, also called a fuel or purchased-power adjustment clause, is updated regularly to pass changes in fuel and purchased-power costs to customers, with a true-up of past over- or under-collection. Higher fuel use and purchase costs therefore reach customers mainly through this charge, faster and with less regulatory review than investment recovered in rate cases. They provide a channel through which the costs, including those for maintaining reliability, may enter retail bills. Because these adjustments also reflect fuel prices and other procurement costs, we interpret them as a broader measure of supply-cost pass-through rather than a measure of reliability costs alone.

2.5 Data

We link data center entry to generator operations, retirement decisions, reliability expenditures and retail cost recovery.

Data centers and zones. Locations, capacity and activation dates come from the Aterio database, which records each building's coordinates, development stage and capacity. We keep 1,999 operating buildings on 1,375 campuses, excluding cryptocurrency mining campuses, whose flexible load differs from that of conventional data centers, buildings with on-site generation, and buildings without an activation date. We measure data center exposure for 98 geographic areas, which we call *zones*: 80 pricing or load zones within the organized markets and 18 balancing authorities outside them. Each building adds its capacity to its zone in the quarter it begins operating. Treatment is defined from this zone-level capacity (Section 3). Every generator, combustion unit and wholesale transaction record in the data below is assigned to the zone in which it is located and inherits that zone's treatment status, while outcomes are measured for each record.

Generation and emissions. EPA's Continuous Emissions Monitoring System (CEMS) reports daily gross generation, heat input, operating hours, and CO_2 and NO_x emissions for fossil combustion units. Our sample contains 5,030 units observed daily from 2014 to 2024. We classify simple-cycle combustion turbines as peakers and all other units (e.g., "Combustion turbine, Combined cycle") as non-peakers, and report results for peakers, non-peakers and all units. EIA Form 923 adds annual generation by technology for all plants. EIA Form 930 provides hourly demand and generation by fuel, from which we construct zone-level measures of system tightness, such as fossil utilization at peak demand and the number of hours with little spare capacity.

Retirements. EIA Form 860 reports each generator's operating status and its planned and actual retirement dates. We build a generator-by-year panel for 2010–2025 from every annual release. A generator retires when it moves to retired status. A retirement is *postponed* when its planned retirement year moves later than in the previous release, and *withdrawn* when a planned retirement disappears while the generator keeps operating; we exclude withdrawals that reflect fuel conversions. We classify gas and oil combustion turbines and internal-combustion engines as peakers, and coal, combined-cycle and steam units as non-peaker fossil generators.

Reliability retention cases. We construct a dataset of reliability retention cases by searching FERC filings under the terminology used in each market. We read each filing to identify the case and locate related ISO, state and utility documents, including reliability studies, agreements and cost-allocation decisions. For each case we record the affected generators, the proposed retirement and retention dates, the compensation, and the customers or areas assigned the cost, distinguishing requested, settled and approved amounts. We also keep cases in which a retirement was reviewed but no agreement was reached.

Reliability payments. FERC's Electric Quarterly Reports (EQR) record each wholesale sale by product, quantity, price and charge. We measure dollars of capacity, uplift, operating reserves, regulation, reactive supply and voltage control, and black start, as well as energy sales. We report three aggregates. *Capacity* is capacity sales. *Ancillary services* is the sum of regulation and frequency response, operating reserves, reactive supply and voltage control, and black start. *Reliability total* is the sum of all reliability products: capacity, ancillary services, and other reliability-related charges (uplift, customer charges and fuel charges). Capacity accounts for most of the reliability total. We place each transaction in a zone using its delivery point or, failing that, its balancing authority or its seller's plant; transactions accounting for 59% of all dollars can be placed. We sum transactions into seller-by-zone records observed in every quarter from 2014Q1 to 2026Q2, with zeros in quarters without a sale. Where a record is tied to a single plant, we classify it by the plant's technology, so that payments to peakers can be separated from payments to other plants. Because not every market has every product (Appendix Table IA1), each product is analyzed only in the markets where it is ever transacted. We also count new contracts by product from the EQR contract filings; only 17% of contract lines can be placed, mainly because contracts delivered at trading hubs carry no local delivery point.[1]

[1]ERCOT does not file the EQR. For Texas we use ERCOT's 60-Day SCED Disclosure Reports, which record each resource's ancillary-service obligations. These measure service commitments rather than pay-

Retail cost recovery. We assemble a panel of fuel and purchased-power adjustment charges from filings and tariff sheets in state utility commission dockets. We prioritize states with high data center capacity relative to electricity demand and cover investor-owned utilities, cooperatives and municipal utilities. For each mechanism we record the adjustment factor, projected and actual fuel and purchased-power costs, the true-up balance and the period covered. We do not substitute average retail prices for missing observations, because those prices combine supply charges with other bill components.

3 Empirical Strategy

We use a stacked difference-in-differences design. For each event cohort g we form a stack s that contains the units with $g_z = g$ (treated) and the never-treated units (clean controls). A stack covers event times $e = t - g \in [-K^-, K^+]$ and enters only if the whole window lies in the data period: $K^- = 8$ and $K^+ = 16$ quarters for EQR and CEMS (2016–2022 and 2016–2020 cohorts), $K^- = 4$ and $K^+ = 8$ years for EIA (2014–2017 cohorts). The stacks are appended and we estimate, by weighted least squares,

$$Y_{ist} \;=\; \sum_{e=-K^-,\, e\neq -1}^{K^+} \beta_e \, \mathbf{1}\{t - g_s = e\} \, D_{is} \;+\; \alpha_{is} \;+\; \lambda_{st} \;+\; \varepsilon_{ist}. \tag{1}$$

Here Y_{ist} is the outcome of record i (generator, CEMS unit or seller × zone record) in stack s at period t (year, day or quarter); $D_{is} = 1$ if the record lies in a treated unit of stack s; α_{is} are record × stack fixed effects, which absorb time-invariant differences between records within a stack; and λ_{st} are stack × period fixed effects, which absorb shocks common to all records of a stack in a period, including seasonality. The coefficients β_e measure the change in the outcome of treated records e periods from activation relative to the period before activation ($e = -1$), net of the change among never-treated records in the same stack. Coefficients with $e < -1$ test for differential pre-trends. [2] Standard errors are clustered at the zone level.

4 Results

We use the event-study design in equation (1) to estimate how generator operations, retirement plans, and wholesale payments change after data center activation. The figures report coefficients relative to the period immediately before activation; the magnitudes below are averages of the post-activation coefficients.

ments and are analyzed separately; the results are reported in Appendix B.

[2] Control observations receive the weight $w_s = \frac{N_s^T/N^T}{N_s^C/N^C}$ and treated observations a weight of one, where N_s^T and N_s^C are the numbers of treated and control zones in stack s, and N^T and N^C are their totals across stacks. These weights give each stack's controls the same relative weight as its treated zones, so that the estimate averages the cohort-specific comparisons in proportion to the number of treated zones. Standard errors are clustered by zone; a zone that serves as a control in several stacks forms a single cluster. Because the number of clusters is modest, we use the CRV1 small-sample correction and compute p-values from a t distribution with $G - 1$ degrees of freedom, where G is the number of zones.

4.1 Generator Operations and Emissions

We first examine how existing generators respond to data center entry. The pre-activation coefficients are generally flat: the fuel-use and CO_2 estimates show no differential pre-trends, although the gross-load event study contains one significant pre-activation coefficient. After activation, a peaking unit in a treated zone generates 33.9 MWh more electricity per day, burns 398 MMBtu more fuel, and operates for 0.31 additional hours (Figure 5). The generation and fuel effects are significant at the 1% level, and the operating-hours effect at the 10% level. We find no significant change among non-peaking units. The increase in generation is therefore concentrated among units that typically operate at the margin.

Greater peaker use also increases emissions. CO_2 emissions rise by 23.7 tons per peaking unit-day, significant at the 1% level (Figure 6). Emissions per MWh do not change. Data center entry therefore raises emissions by increasing peaker output, not by changing the emissions intensity of generation.

4.2 Retirement Plans

We next examine whether data center entry changes generator retirement plans. The event-study coefficients show no differential pre-trends. After activation, the probability that a planned retirement is postponed rises by 0.28 percentage points for all generators and by 0.52 percentage points for fossil generators (Figure 3). The effects are significant at the 5% and 1% levels, respectively. Postponements increase for both peakers and non-peaker fossil generators, by 0.44 and 0.66 percentage points, and postponed peaker capacity rises by 0.11 MW per generator-year. Each estimate is significant at the 5% level or better.

The result is similar when we broaden the outcome to include retirement plans that disappear while the generator remains in operation (Figure 4). Under this measure, delayed retirement rises by 0.80 percentage points for fossil generators. The effect is largest for non-peaker fossil units, including coal and steam plants: their probability of delay rises by 1.19 percentage points and their delayed capacity by 2.12 MW per generator-year. All three estimates are significant at the 5% level.

We find no significant change in completed retirements during the sample period. The observed response occurs in retirement plans: generators scheduled to close are instead kept available for longer, while many of their revised retirement dates fall beyond the period we observe.

4.3 Energy and Capacity Payments

We then examine whether these physical responses are reflected in wholesale transactions. Energy revenues, quantities, and capacity revenues show no differential pre-trends. After data center activation, fossil sellers receive an additional $3.774 million in energy-sale revenues and sell 122,939 MWh more electricity per seller-zone record and quarter (Figure 7). Energy revenues increase significantly at the 1% level and quantities at the 5% level.

Capacity-sale revenues also rise, by $359 thousand per record and quarter, significant at the 5% level (Figure 8).

The number of contracts increases as well. Fossil sellers hold 0.29 more active contracts per record and quarter, including 0.27 more capacity contracts and 0.31 more contracts for reliability products (Figure 9). All three estimates are significant at the 5% level. Because only 17% of contract lines can be assigned to a zone, we interpret these contract results as supporting evidence.

4.4 Ancillary Services

We next examine payments for reliability services other than capacity. The ancillary-service estimates for fossil sellers exhibit differential pre-trends and are therefore not interpreted causally. Estimated post-activation effects on total ancillary-service sales and on the sum of all reliability products are positive but not statistically significant (Figure 8).

The reactive-power event studies also each contain one significant pre-activation coefficient, so we interpret their post-activation estimates as suggestive. Because Order No. 904 changed the treatment of reactive-power charges beginning in 2025, the reactive-power sample ends in 2024Q4. After activation, reactive supply and voltage-control revenues increase by $33.6 thousand per record for fossil sellers and by $3.11 thousand across all sellers; both estimates are significant at the 5% level.

4.5 Weather Anomalies

Finally, we test whether unusual weather amplifies the effects of data center entry. We interact treatment with a standardized, zone-specific weather anomaly based on heating and cooling degree days; one standard deviation is approximately 167 degree days. In the annual specification, the reported interactions show no differential pre-trends. Quarterly specifications reveal pre-activation divergence for several outcomes, however, so we interpret the weather estimates as suggestive. The anomaly is driven mainly by heating degree days, and a separate specification finds no amplification during hot extremes. The event-study coefficients reported in Figures 10 and 11 are binned by year.

For fossil sellers, a one-standard-deviation weather anomaly increases the post-activation effect on energy sales by an additional 47,034 MWh per record and quarter. It also increases reactive-power payments by $8.3 thousand and ancillary-service payments by $16.3 thousand. Across all sellers, energy revenues rise by an additional $938 thousand. The reactive-power and all-seller energy-revenue interactions are significant at the 1% level; the fossil energy-quantity and ancillary-service interactions are significant at the 5% level. The results therefore point to amplification under cold-weather stress rather than a general response to extreme temperatures.

Overall, data center entry increases peaker generation and emissions, delays fossil-generator retirements, and raises energy and capacity payments to incumbent generators. Some of these effects become larger when weather places additional stress on the system.

5 To-dos

5.1 Retention Cases

The retirement-plan estimates show that generators remain available for longer after data center activation, but they do not reveal when this response is required by a system operator or who bears the resulting cost. We will use the retention-case dataset to study this institutional margin directly. The analysis will document how often retirement reviews lead to retention agreements in zones with and without data center growth, estimate whether data center exposure increases their probability and cost, and use the cost-allocation decision in each case to measure how much of the expense falls on customers beyond the generator's zone. Because these agreements are not that often, we will present the regression estimates together with case-level evidence.

5.2 Utility Rates

The preceding results establish that data center activation raises generation by peaking units, delays planned retirements, and increases some wholesale energy and reliability payments. The final step is to determine whether these higher supply costs reach retail customers between rate cases. Using the panel of fuel and purchased-power adjustment charges, we will estimate event studies analogous to equation (1) for three outcomes: the utility's realized fuel and purchased-power cost per unit sold, the adjustment factor charged to customers, and the true-up balance. Taken together, these outcomes distinguish higher procurement costs from contemporaneous pass-through and costs deferred to future bills. We will estimate these responses separately for investor-owned utilities, cooperatives, and municipal utilities.

6 Conclusion

The cost of serving a large new load is not only the cost of the energy it consumes. A reliable power system must also hold enough capacity, reserves, and voltage support, in the right places and at the right times, to keep supply and demand balanced under stress. When a large, concentrated load arrives faster than new generation and transmission can be built, these requirements are met by drawing on, and paying to maintain, the existing system. These reliability costs are not totally included in wholesale energy prices and are often recovered through charges that are spread across many customers. Understanding the reliability cost is necessary both to measure the full cost of data center growth and to understand who are bearing it.

We find that data center load draws on the system's reliability margin. It is met largely by peaking units, which normally stand ready for peaks and contingencies: after activation, a peaking unit generates 33.9 MWh more per day and emits 23.7 more tons of CO_2, while other fossil units show no significant change. The system also keeps existing capacity available for

longer: the annual probability that a fossil generator postpones its planned retirement rises by 0.52 percentage points. Wholesale payments for availability rise accordingly, as fossil sellers' capacity revenues increase by $359 thousand per seller and quarter and they hold more capacity contracts. Payments for voltage support also rise.

Debates over large-load tariffs and interconnection rules have focused on the cost of grid connections and new infrastructure. Our results suggest that the cost of keeping the existing system reliable belongs in the same discussion. These costs arise after a data center connects, through the operation and retention of existing generators and through payments for availability and reliability services, and they are typically recovered from all customers in a zone rather than from the load that triggered them. In addition, the reliance on the existing fossil fleet, and particularly on peaking units, has environmental consequences. If new load is met in the short run by older and less efficient units, and keeps some of them from retiring, the emissions cost of data center growth depends on how quickly new clean capacity and transmission can follow.

References

Callaway, D. S., Fowlie, M., and McCormick, G. (2018). Location, location, location: The variable value of renewable energy and demand-side efficiency resources. *Journal of the Association of Environmental and Resource Economists*, 5(1):39–75.

Cicala, S. (2015). When does regulation distort costs? Lessons from fuel procurement in US electricity generation. *American Economic Review*, 105(1):411–444.

Darmouni, O., Lehner, C., and Zhang, Y. (2026). The Energy Cost of AI and Data Centers.

Davis, L. and Hausman, C. (2016). Market impacts of a nuclear power plant closure. *American Economic Journal: Applied Economics*, 8(2):92–122.

Du, X. and Guo, J. (2026). Grid Lock-in and the Development of AI.

Fabra, N. and Reguant, M. (2014). Pass-through of emissions costs in electricity markets. *American Economic Review*, 104(9):2872–2899.

Fabrizio, K. R., Rose, N. L., and Wolfram, C. D. (2007). Do markets reduce costs? Assessing the impact of regulatory restructuring on US electric generation efficiency. *American Economic Review*, 97(4):1250–1277.

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

Heiss, A., Sautner, Z., and Schmid, T. (2026). Powering AI: How Do Data Centers Affect Renewable Energy Investment?

Holland, S. P. and Mansur, E. T. (2008). Is real-time pricing green? The environmental impacts of electricity demand variance. *Review of Economics and Statistics*, 90(3):550–561.

Joskow, P. and Tirole, J. (2007). Reliability and competitive electricity markets. *RAND Journal of Economics*, 38(1):60–84.

Mamkhezri, J., Sun, X., and Yang, Y. (2026a). The Hidden Cost of the Cloud: Data Centers and Transmission-Induced Electricity Prices.

Mamkhezri, J., Sun, X., and Yang, Y. (2026b). Site and Scale: Data Centers, Grid Congestion, and Welfare.

Meeks, R., Pless, J., Qi, Z., and Wang, Z. (2026). Who Pays for Growth? Evidence from Data Centers and the Grid.

North American Electric Reliability Corporation (2026). Reliability insights: Large computational load risks due to voltage sensitivity. Reliability insights, North American Electric Reliability Corporation.

PJM Interconnection (2026). Powering reliability through market design: Addressing rising demand and constrained supply, and stimulating investment to support durable reliability. White paper, PJM Interconnection.

Reguant, M. (2014). Complementary bidding mechanisms and startup costs in electricity markets. *Review of Economic Studies*, 81(4):1708–1742.

Scalera, C., Bosetti, V., and Pecci, F. (2026). Large-Load Electricity Customers and Spillovers on Household Prices: Evidence from U.S. Data Centers. Number: DP21854.

Figures and Tables

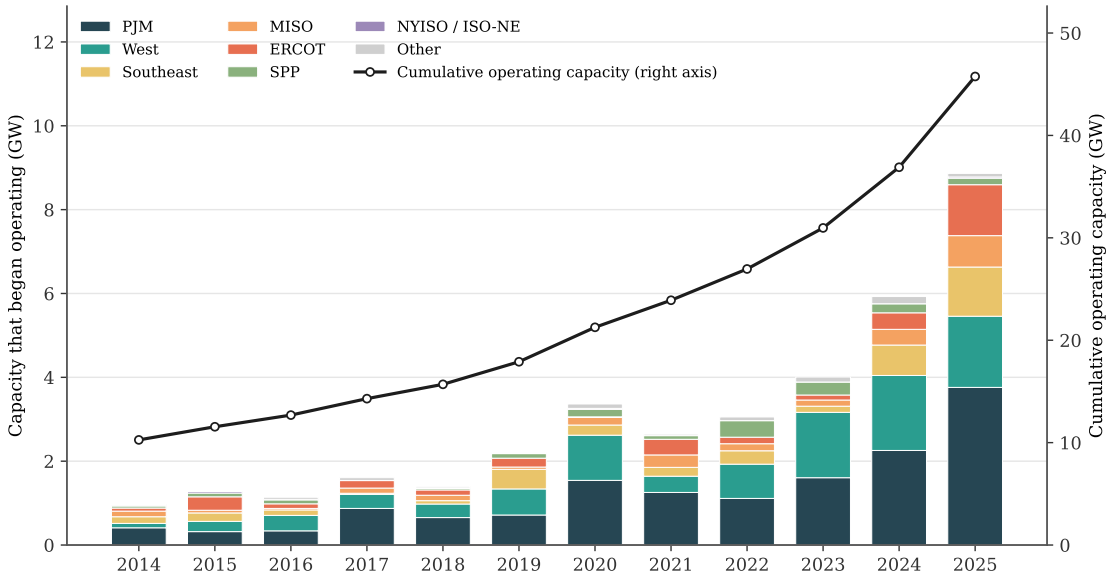


Figure 1: Data center capacity in the analysis sample

Notes: Bars show the capacity (GW) of data center buildings that began operating in each year, by wholesale market. The line shows cumulative operating capacity, including buildings that began operating before 2014. The sample includes operating buildings in the contiguous United States that began operating by the end of 2025 and excludes campuses with cryptocurrency mining. "West" combines CAISO and the Western Energy Imbalance Market.

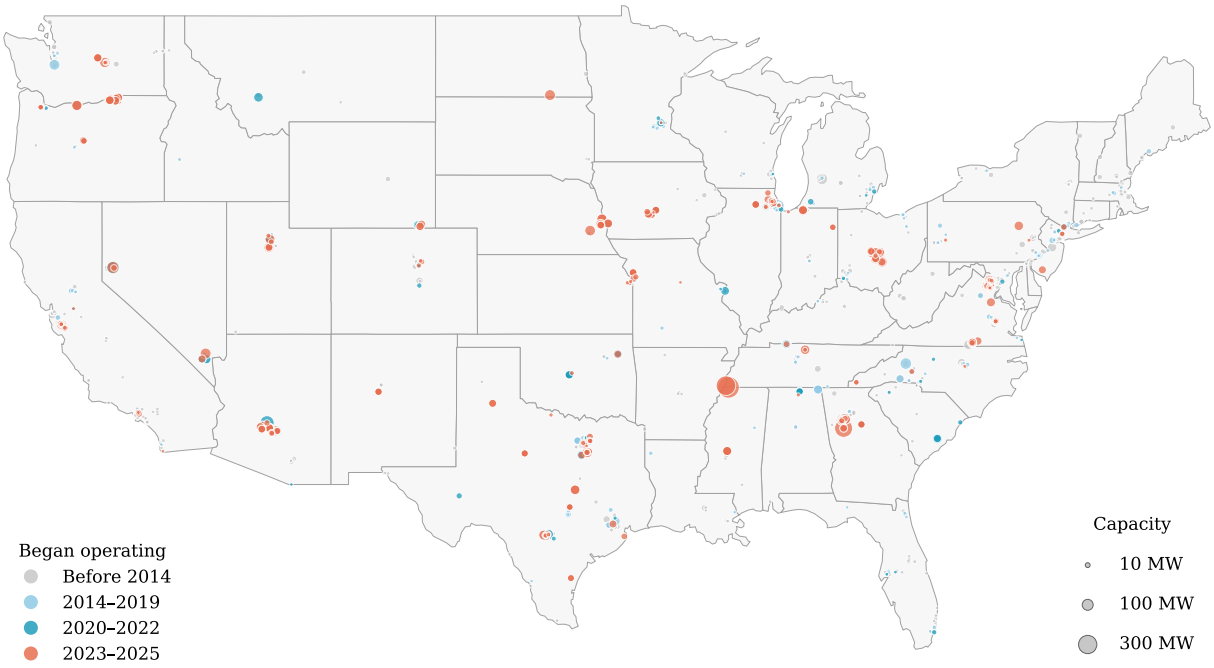
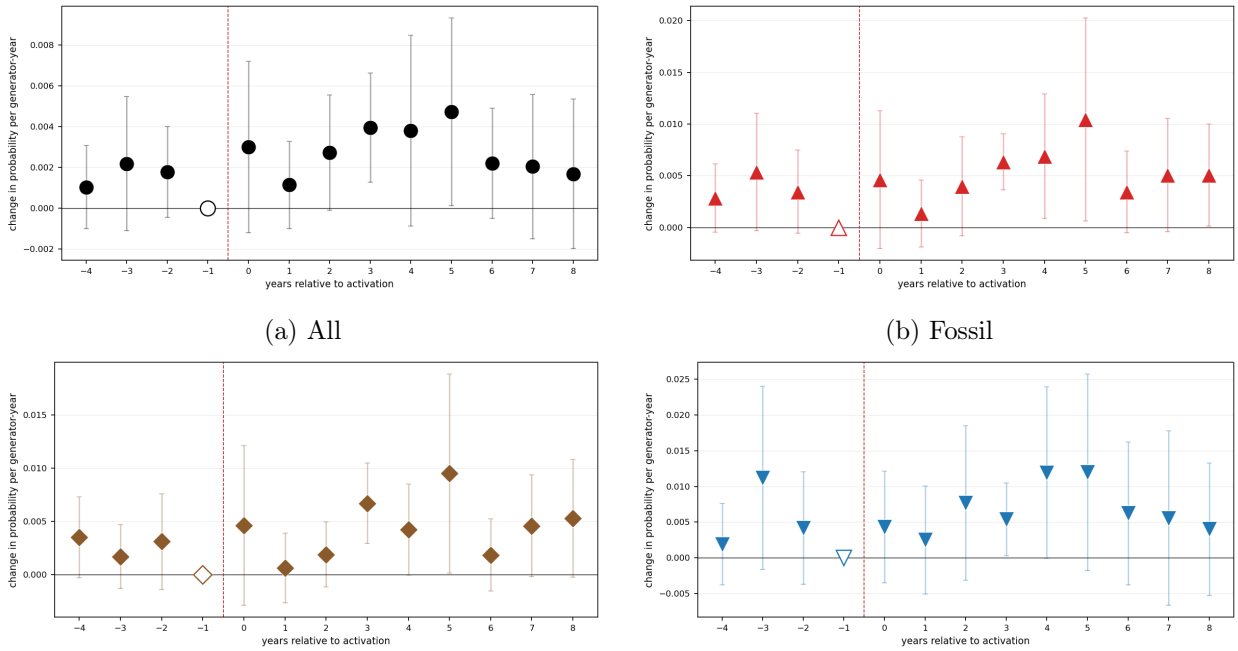


Figure 2: Location of data centers in the analysis sample

Notes: Each circle is an operating data center building. Circle size is proportional to capacity (MW), and color indicates when the building began operating. Same sample as Figure 1.



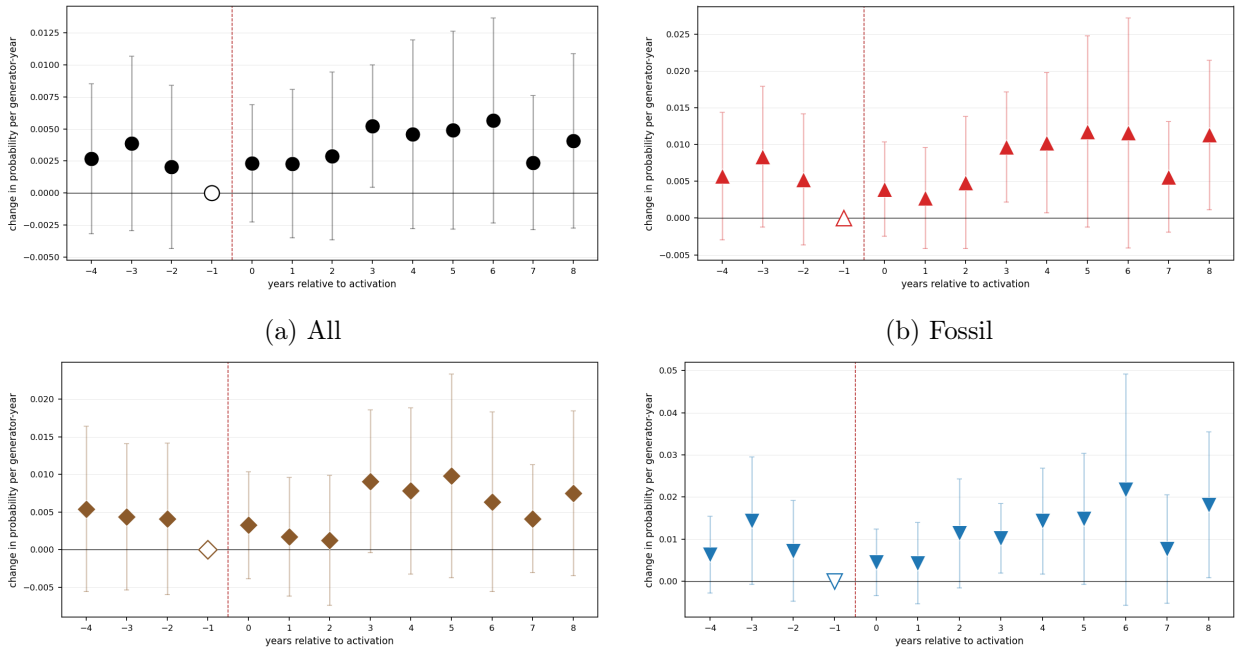
(a) All

(b) Fossil

(c) Peaker

(d) Non-peaker

Figure 3: **Event-study estimates: postponed retirement plans (EIA generator × year).** Notes: The figure presents event-study estimates for EIA-860 generators operating on 1 January of each year, 2010–2025. The dependent variable is an indicator equal to one if the generator's planned retirement year in the year's EIA-860 is later than the most recent earlier plan; coefficients are changes in the annual probability. Panels: (a) all generators, (b) fossil generators, (c) peakers (gas and oil combustion turbines and internal-combustion engines), (d) non-peaker fossil generators (coal, gas combined cycle, gas and oil steam). Event time is in years, $e \in [-4, 8]$, which admits the units activating in 2014–2017. Event time is measured relative to the activation quarter g of the record's unit, and $e = -1$ is omitted. Regressions include record × stack and stack × period fixed effects; controls are never-treated units. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.



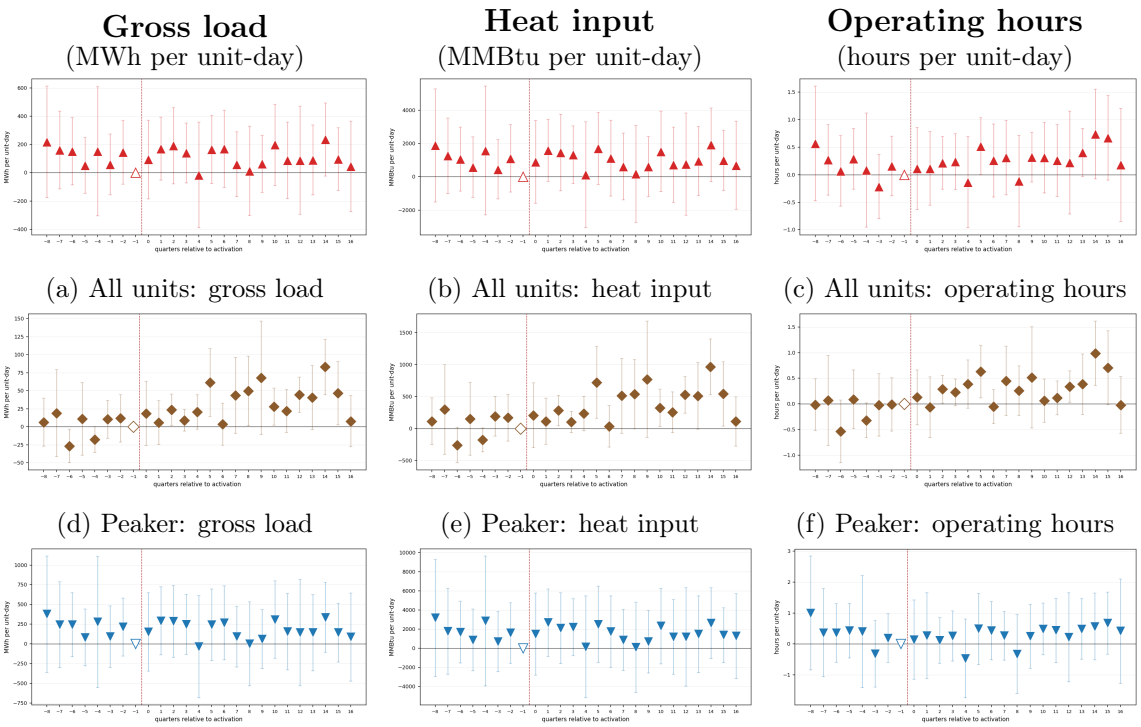
(a) All

(b) Fossil

(c) Peaker

(d) Non-peaker

Figure 4: **Event-study estimates: delayed retirement plans (EIA generator × year).** Notes: The figure presents event-study estimates for EIA-860 generators operating on 1 January of each year, 2010–2025. The dependent variable is an indicator equal to one if the plan is postponed or withdrawn (a planned retirement year dropped while the generator keeps operating; fuel conversions excluded); coefficients are changes in the annual probability. Panels: (a) all generators, (b) fossil generators, (c) peakers (gas and oil combustion turbines and internal-combustion engines), (d) non-peaker fossil generators (coal, gas combined cycle, gas and oil steam). Event time is in years, $e \in [-4, 8]$, which admits the units activating in 2014–2017. Event time is measured relative to the activation quarter g of the record's unit, and $e = -1$ is omitted. Regressions include record × stack and stack × period fixed effects; controls are never-treated units. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.

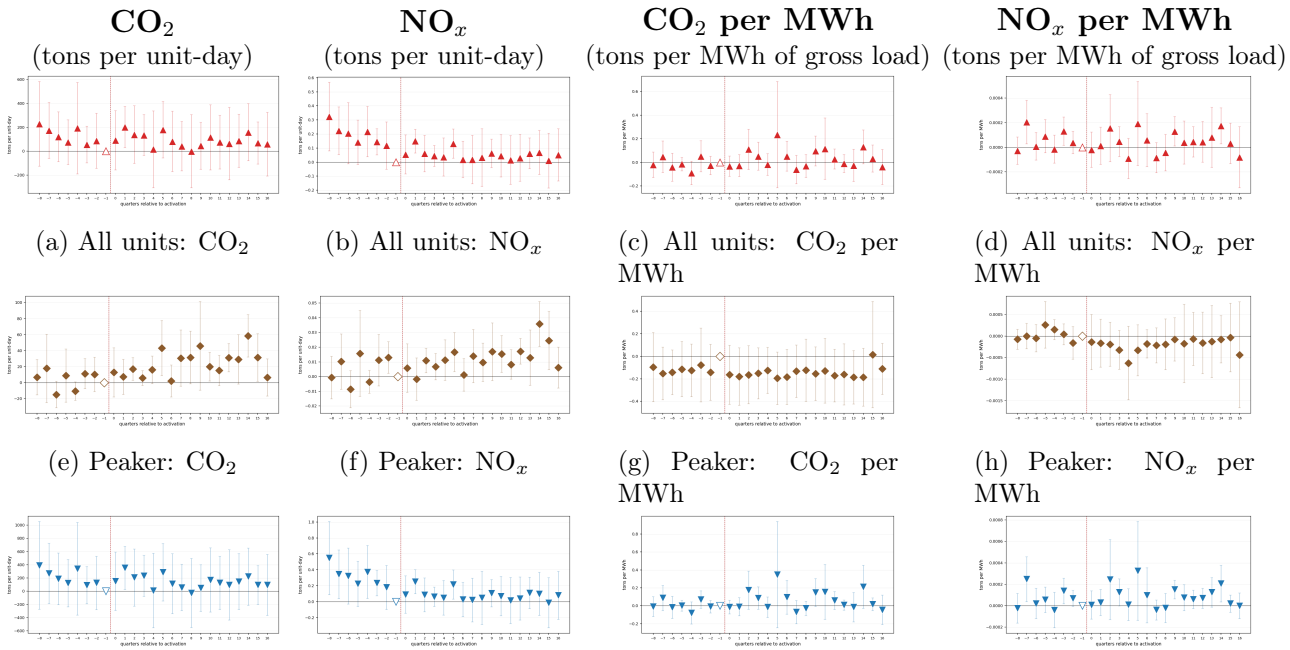


(a) All units: gross load (b) All units: heat input (c) All units: operating hours

(d) Peaker: gross load (e) Peaker: heat input (f) Peaker: operating hours

(g) Non-peaker: gross load (h) Non-peaker: heat input (i) Non-peaker: operating hours

Figure 5: **Event-study estimates: gross load, heat input and operating hours (CEMS unit × day).** Notes: The figure presents event-study estimates for CEMS units, 2014-01-01 to 2024-12-31. Columns: gross load (MWh per unit-day), heat input (MMBtu per unit-day), operating hours (hours per unit-day), in levels. Rows: all CEMS units (combustion units reporting to CEMS), peakers (simple-cycle combustion turbines), non-peakers. Event time is in quarters (each day carries the quarter of its date), $e \in [-8, 16]$, which admits the units activating in 2016–2020. Event time is measured relative to the activation quarter g of the record's unit, and $e = -1$ is omitted. Regressions include record × stack and stack × period fixed effects; controls are never-treated units. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.

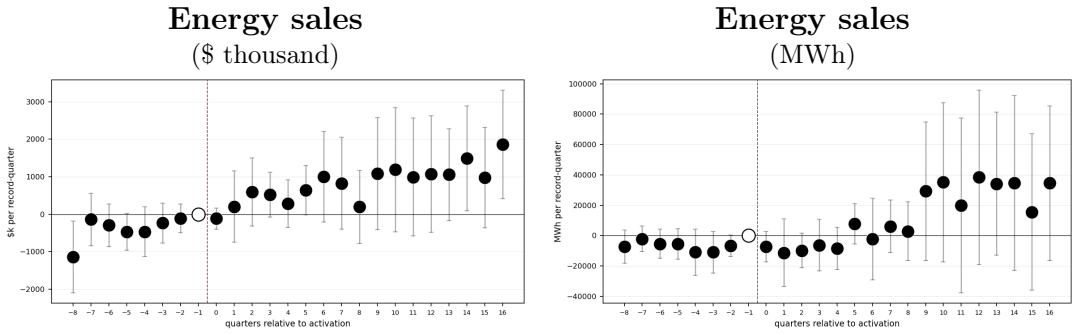


(a) All units: CO_2 (b) All units: NO_x (c) All units: CO_2 per MWh (d) All units: NO_x per MWh

(e) Peaker: CO_2 (f) Peaker: NO_x (g) Peaker: CO_2 per MWh (h) Peaker: NO_x per MWh

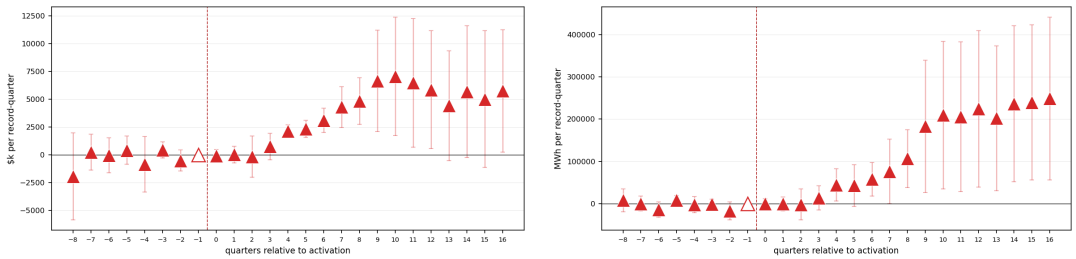
(i) Non-peaker: CO_2 (j) Non-peaker: NO_x (k) Non-peaker: CO_2 per MWh (l) Non-peaker: NO_x per MWh

Figure 6: **Event-study estimates: CO_2 and NO_x emissions, totals and per MWh of gross load (CEMS unit $\times$ day).** Notes: The figure presents event-study estimates for CEMS units, 2014-01-01 to 2024-12-31. Columns: CO_2 (tons per unit-day), NO_x (tons per unit-day), CO_2 per MWh (tons per MWh of gross load), NO_x per MWh (tons per MWh of gross load), in levels; the per-MWh outcomes are the day's emissions divided by its gross load, defined only on days with positive gross load (days with zero or no reported gross load are excluded). Rows: all CEMS units (combustion units reporting to CEMS), peakers (simple-cycle combustion turbines), non-peakers. Event time is in quarters (each day carries the quarter of its date), $e \in [-8, 16]$, which admits the units activating in 2016–2020. Event time is measured relative to the activation quarter g of the record's unit, and $e = -1$ is omitted. Regressions include record $\times$ stack and stack $\times$ period fixed effects; controls are never-treated units. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.



(a) All: energy sales

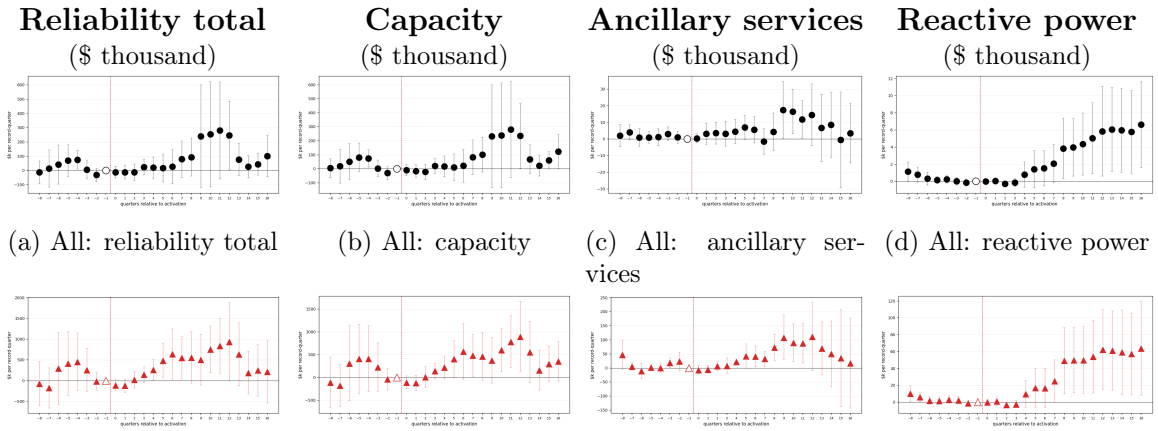
(b) All: energy sales



(c) Fossil: energy sales

(d) Fossil: energy sales

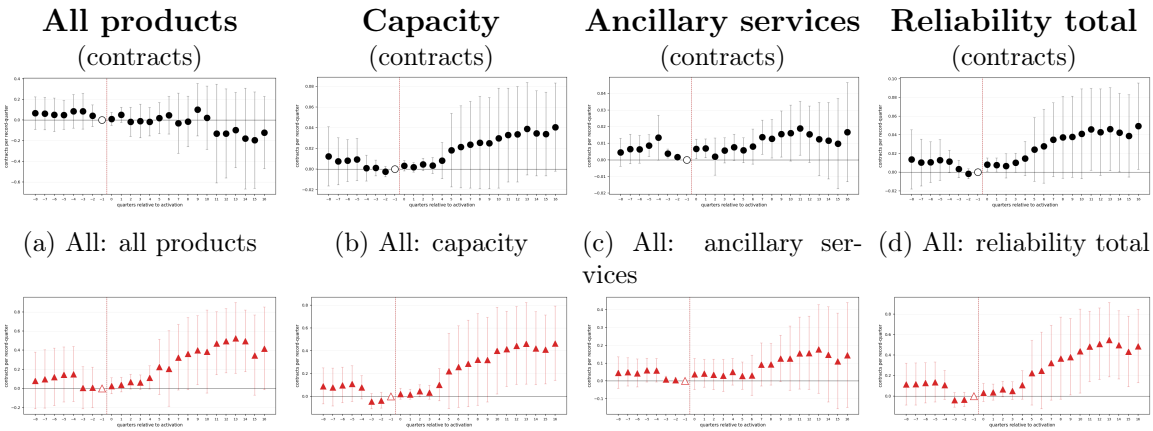
Figure 7: **Event-study estimates: energy sales (EQR seller × zone).** Notes: The figure presents event-study estimates for EQR seller × zone records, 2014Q1–2026Q2. Columns: energy sales in $ thousand and in MWh, per record and quarter, in levels, zero in quarters without a sale. Each outcome is estimated only in markets that carry the product (energy sales: CAISO, ERCOT, MISO, NEISO, NYISO, PJM, SPP, WECC). Rows: all records; records whose delivery points all link to a single fossil plant. Event time is in quarters, $e \in [-8, 16]$, which admits the units activating in 2016–2022. Event time is measured relative to the activation quarter g of the record's unit, and $e = -1$ is omitted. Regressions include record × stack and stack × period fixed effects; controls are never-treated units. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.



(a) All: reliability total (b) All: capacity (c) All: ancillary services (d) All: reactive power

(e) Fossil: reliability total (f) Fossil: capacity (g) Fossil: ancillary services (h) Fossil: reactive power

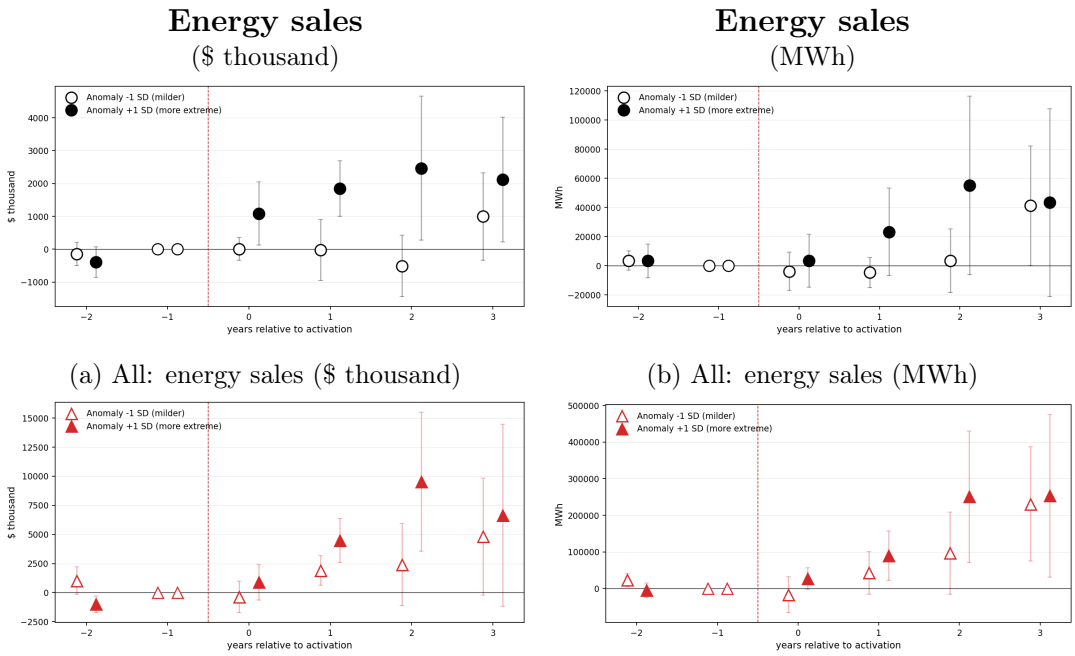
Figure 8: **Event-study estimates: reliability products (EQR seller × zone).** Notes: The figure presents event-study estimates for EQR seller × zone records, 2014Q1–2026Q2. Columns: reliability total, the sum of all reliability products (capacity, ancillary services, and other charges: uplift, customer charges and fuel charges); capacity sales; ancillary-service sales, the sum of regulation and frequency response, operating reserves, reactive supply and voltage control, and black start; and reactive-power sales, a component of ancillary services; all in $ thousand, per record and quarter, in levels, zero in quarters without a sale. Each outcome is estimated only in markets that carry the product (reliability: CAISO, ERCOT, MISO, NEISO, NYISO, PJM, SPP, WECC; capacity: CAISO, ERCOT, MISO, NEISO, NYISO, PJM, SPP, WECC; ancillary services: CAISO, ERCOT, MISO, NEISO, NYISO, PJM, SPP, WECC; reactive power: CAISO, CAR, FLA, MISO, NEISO, NYISO, PJM, SE, SPP, WECC). Rows: all records; records whose delivery points all link to a single fossil plant. Event time is in quarters, $e \in [-8, 16]$, which admits the units activating in 2016–2022. Event time is measured relative to the activation quarter g of the record's unit, and $e = -1$ is omitted. Regressions include record × stack and stack × period fixed effects; controls are never-treated units. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.



(a) All: all products (b) All: capacity (c) All: ancillary services (d) All: reliability total

(e) Fossil: all products (f) Fossil: capacity (g) Fossil: ancillary services (h) Fossil: reliability total

Figure 9: **Event-study estimates: active contracts (EQR seller × zone).** Notes: The figure presents event-study estimates for the number of active contracts per EQR seller × zone record and quarter, in levels, zero where the record has no such contract. An active contract has a product line of the column's type filed in the quarter. Columns: all products (all markets); capacity; ancillary services; reliability total (capacity, ancillary services, and must-run). Product columns are estimated only in markets that carry the product (capacity: CAISO, ERCOT, MISO, NEISO, NYISO, PJM, SPP, WECC; ancillary services: CAISO, ERCOT, MISO, NEISO, NYISO, PJM, SPP, WECC; reliability products: CAISO, ERCOT, MISO, NEISO, NYISO, PJM, SPP, WECC). Only 17.4% of contract lines can be placed in a zone, so the counts cover 54,866 of 314,815 contracts. Rows: all records; fossil records. Event time is in quarters, $e \in [-8, 16]$. Event time is measured relative to the activation quarter g of the record's unit, and $e = -1$ is omitted. Regressions include record × stack and stack × period fixed effects; controls are never-treated units. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.



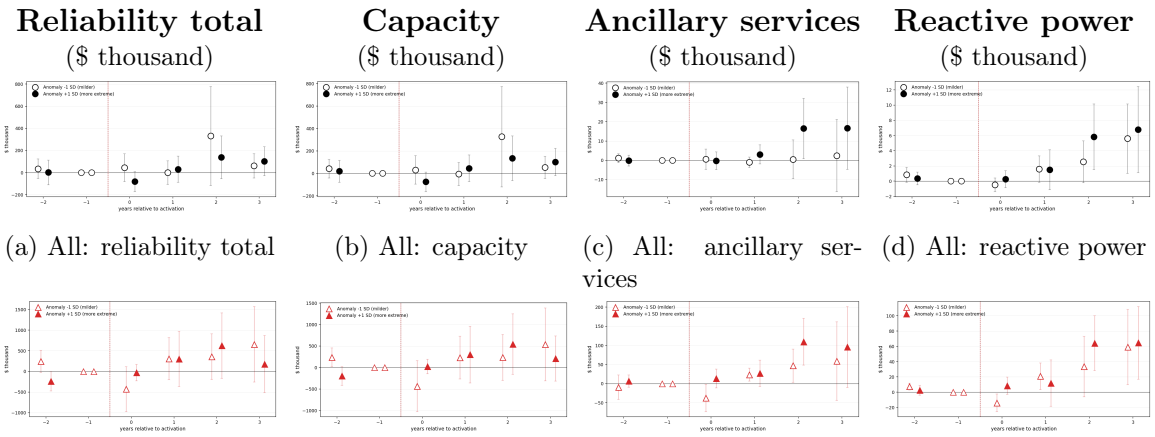
(a) All: energy sales ($ thousand)

(b) All: energy sales (MWh)

(c) Fossil: energy sales ($ thousand)

(d) Fossil: energy sales (MWh)

Figure 10: **Weather-anomaly event study: energy sales (EQR seller × zone).** Notes: The figure presents weather-anomaly event-study estimates for EQR seller × zone records, 2014Q1–2025Q4, in levels, zero in quarters without a sale; each outcome is estimated only in markets that carry the product. Event time is binned into years relative to the activation quarter, years −2 to 3 (window −8 to +15 quarters, 2014Q1–2025Q4, the years with weather data; year −1 = the four quarters before activation, omitted). Hollow markers: the path at a degree-day anomaly one standard deviation below the zone's normal for the quarter; filled markers: one standard deviation above. The anomaly is the zone-quarter's heating plus cooling degree days minus the zone's mean for that quarter of the year (2014–2025), standardised. The regression interacts each event-year indicator with the anomaly and includes the anomaly, treated × anomaly and post × anomaly, record × stack and stack × quarter fixed effects; controls are never-treated units. Rows: all records; records whose delivery points all link to a single fossil plant. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.



(a) All: reliability total (b) All: capacity (c) All: ancillary services (d) All: reactive power

(e) Fossil: reliability total (f) Fossil: capacity (g) Fossil: ancillary services (h) Fossil: reactive power

Figure 11: **Weather-anomaly event study: reliability products (EQR seller × zone).** Notes: The figure presents weather-anomaly event-study estimates for EQR seller × zone records, 2014Q1–2025Q4, in levels, zero in quarters without a sale; each outcome is estimated only in markets that carry the product. Columns: reliability total (the sum of capacity, ancillary services, and other charges: uplift, customer charges and fuel charges), capacity, ancillary services (regulation and frequency response, operating reserves, reactive supply and voltage control, black start) and reactive power (a component of ancillary services). Event time is binned into years relative to the activation quarter, years −2 to 3 (window −8 to +15 quarters, 2014Q1–2025Q4, the years with weather data; year −1 = the four quarters before activation, omitted). Hollow markers: the path at a degree-day anomaly one standard deviation below the zone's normal for the quarter; filled markers: one standard deviation above. The anomaly is the zone-quarter's heating plus cooling degree days minus the zone's mean for that quarter of the year (2014–2025), standardised. The regression interacts each event-year indicator with the anomaly and includes the anomaly, treated × anomaly and post × anomaly, record × stack and stack × quarter fixed effects; controls are never-treated units. Rows: all records; records whose delivery points all link to a single fossil plant. Standard errors are clustered by unit (zone or balancing authority), and whiskers are 95% confidence intervals.

Internet Appendix

A Reliability mechanisms in detail

Keeping the grid reliable requires more than producing enough energy each hour. System operators rely on a set of mechanisms, each paying for a different dimension of reliability. Most of these payments are separate from the energy price, so a rise in reliability costs can go unnoticed if one looks only at wholesale or retail energy prices.

Table IA1 summarizes which mechanisms exist in each market and how each is paid for.

Table IA1: Reliability mechanisms by market

	PJM	MISO	SPP	NYISO	ISO-NE	CAISO	ERCOT	Non-RTO
Capacity	M	M	O	M	M	O	–	U
Operating reserves	M	M	M	M	M	M	M[a]	T
Regulation	M	M	M	M	M	M	M[a]	T
Reactive supply and voltage control	T[b]	T[b]	T[b]	T[b]	T[b]	–[c]	–[a]	T[b]
Uplift (make-whole)	Y	Y	Y	Y	Y	Y	Y[a]	–
Black start	T	T	T	T	T	T	T[a]	T
Retention agreements	RMR	SSR	SSR[d]	RSSA	RMR/COS	RMR	RMR[a]	–

Notes: M = procured through an ISO market. T = paid under a cost-based tariff rate, typically the open-access transmission tariff schedules (Schedule 2 for reactive power, 3 for regulation, 5–6 for reserves) or an ISO tariff schedule. O = a resource adequacy obligation met bilaterally, with no centralized capacity auction. U = met through utility resource planning. Y = make-whole payments exist. – = not procured or not separately paid. RMR = reliability-must-run; SSR = System Support Resource; RSSA = Reliability Support Services Agreement; COS = cost-of-service agreement. Non-RTO covers areas outside the organized markets, mainly the Southeast and the West outside CAISO. [a] ERCOT is not FERC-jurisdictional, and its transactions do not appear in the EQR; ERCOT ancillary services are measured from ERCOT's own data. ERCOT requires reactive capability as a condition of interconnection without separate payment. [b] Until 2025. FERC Order No. 904 (October 2024, effective January 2025, with later effective dates of up to mid-2025 allowed for PJM, NYISO and ISO-NE) ended payment for reactive power within the standard power factor range; payment remains when the operator directs a generator outside that range. MISO had ended such payments in its own tariff in December 2022; in September 2025 the D.C. Circuit vacated FERC's approval of that change on the ground that it took effect immediately (*Capital Power Corp. v. FERC*, No. 23-1134). [c] CAISO's tariff requires generators to operate within a specified power factor range without compensation; generators are paid only when directed outside that range, based on forgone energy sales. [d] SPP's System Support Resource program was approved by FERC in March 2026; before then SPP had a retirement study process but no retention agreement mechanism. For each product, the analysis keeps only markets in which the product is transacted (Section 2).

Capacity. Capacity payments compensate generators for being available to produce when needed, whether or not they actually run. They address *resource adequacy*: whether the system has enough installed generation to meet peak demand plus a safety margin. In PJM, ISO-NE, NYISO and MISO, load-serving entities must procure capacity, largely

through centralized capacity auctions. Vertically integrated utilities meet the same requirement through their own planning. SPP and CAISO impose resource adequacy obligations but have no centralized capacity auction; load-serving entities meet them bilaterally. ERCOT has no capacity market and relies on scarcity pricing in the energy market instead. Because data centers add load at all hours, including the system peak, they raise the capacity the system must hold.

Operating reserves. Operating reserves are generating capacity held back, unloaded, so that it can respond within minutes if a large generator or transmission line fails. Spinning (synchronized) reserves come from units already running and connected to the grid. Non-spinning or supplemental reserves come from units that can start and ramp up quickly. ERCOT procures several reserve products, including Responsive Reserve Service (RRS), which responds within seconds to minutes to arrest a drop in frequency, ERCOT Contingency Reserve Service (ECRS) and Non-Spinning Reserve. Reserve requirements are typically set relative to the largest single contingency and to the uncertainty in load, so large and concentrated new loads can raise them. All the ISOs procure reserves through markets. Outside the organized markets, transmission providers supply reserves under cost-based rates in their open-access transmission tariffs (Schedules 5 and 6).

Regulation. Regulation is the continuous, second-by-second adjustment of generator output to keep supply and demand exactly balanced and system frequency at its target. It is provided by units under automatic generation control. All the ISOs procure regulation through markets; outside them it is supplied under a cost-based tariff rate (Schedule 3).

Reactive power and voltage control. Reactive power and voltage control are two sides of one service: voltage control is the objective, and supplying or absorbing reactive power is the means. FERC treats them as a single ancillary service, and they appear as a single product in the Electric Quarterly Reports. Electricity in an alternating-current system has two components. *Real power*, measured in megawatts, does useful work. *Reactive power*, measured in megavars, does no useful work itself but is needed to maintain voltage at the levels at which equipment is designed to operate. Transmission lines consume more reactive power as they carry more current, so voltage tends to sag in heavily loaded areas unless reactive power is supplied nearby. Unlike real power, reactive power cannot be transported efficiently over long distances; it must be produced close to where it is needed. It is supplied mainly by generators, which can raise or lower their reactive output by adjusting their excitation, and by network devices such as capacitor banks and static compensators. Under FERC's open-access transmission tariffs, generators are paid for this service under *Reactive Supply and Voltage Control from Generation Sources Service* (Schedule 2), usually through a cost-based annual revenue requirement for each plant approved by FERC. This is the case in PJM, MISO, SPP, NYISO, ISO-NE and outside the organized markets. CAISO does not pay generators for reactive power within the required power factor range, and ERCOT requires reactive capability as a condition of interconnection without separate payment. FERC Order No. 904, issued in October 2024 and effective in 2025, ended payment for reactive power within the standard power factor range in all FERC-jurisdictional markets; payment remains for operation outside that range when the operator directs it. Reactive power payments in the EQR therefore fall mechanically from 2025, which our estimates must

account for.

Voltage support is a particularly informative signal of data center demand. Data centers are dense concentrations of power electronics, such as server power supplies and uninterruptible power supply (UPS) systems, that are highly sensitive to voltage quality. When voltage deviates from its normal range even briefly, these systems are designed to protect the servers by disconnecting from the grid and switching to backup power. Most residential and conventional industrial loads, by contrast, tolerate short voltage dips. As a result, a cluster of data centers places unusual demands on local voltage control: the system must hold voltage within a tighter band, and it must also be prepared for the sudden loss of hundreds of megawatts of load if a disturbance causes many facilities to disconnect at once. Grid reliability authorities have documented such events in data center regions (North American Electric Reliability Corporation, 2026). Because reactive power must be supplied locally, these needs fall on generators near data center clusters. They can raise reactive power payments and make it harder to let nearby generators retire.

Black start. Black-start units can restart without power from the grid and are used to restore the system after a blackout. They are procured under long-term arrangements and have little reason to respond to short-run load growth. Every market procures black-start service, usually under a cost-based tariff schedule.

Uplift. Market prices do not always cover the cost of running the units the operator needs. A unit may be committed for reliability, for example to support voltage or relieve a transmission constraint, even though the energy price is below its operating cost. The operator then pays the difference through *make-whole* or *uplift* payments, which are charged to market participants outside the energy price. As load grows faster than supply, operators take more of these manual, out-of-market actions, which raises production costs and shifts costs across customers (PJM Interconnection, 2026). Uplift exists only where an operator runs a centralized market and commits units centrally, that is, in the ISOs. Its names vary: balancing operating reserve credits in PJM, revenue sufficiency guarantee and make-whole payments in MISO and SPP, bid cost recovery in CAISO, and net commitment period compensation in ISO-NE. Outside the organized markets, vertically integrated utilities absorb these costs in their own operations, so no separate uplift payment exists.

Reliability retention agreements. The most direct form of out-of-market reliability procurement is an agreement that keeps a retiring generator online. When an owner notifies the system operator that it intends to retire or suspend a unit, the operator studies whether the grid can still meet reliability standards without it. If not, typically because the plant supports voltage or relieves a transmission constraint in a load pocket, the operator can require the unit to stay available until transmission upgrades or replacement capacity are in place. In exchange, the owner receives cost-based compensation under an agreement approved by FERC, and the cost is recovered from load. These agreements go by different names in different markets: reliability-must-run (RMR) agreements in PJM, CAISO and ISO-NE, where they are also called cost-of-service agreements; System Support Resource (SSR) agreements in MISO; and Reliability Support Services Agreements (RSSA) in NYISO. ERCOT has its own RMR process under Texas rules. Reliability retention agreements. Outside the organized markets, a vertically integrated utility that needs a plant for reliability

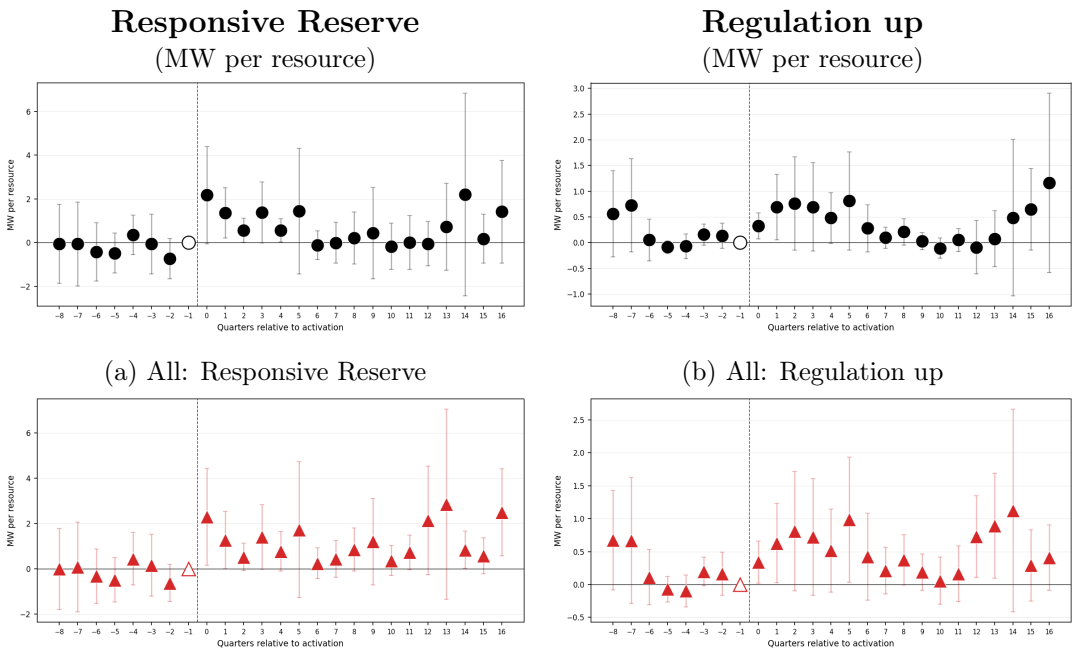
simply keeps it in service, so no separate agreement is filed. We refer to all of them as *reliability retention agreements*.

How the cost of these agreements is allocated differs across markets and cases. In some cases the cost is spread across a whole zone or region. In others it is assigned in proportion to the load each area would have had to shed without the unit, so that the cost falls on the customers who benefit most. Reliability retention agreements combine two of our margins: they postpone retirements, and they create reliability payments whose allocation determines who bears the cost. Because data center load is concentrated in specific transmission zones, it can raise the number of retirements found to threaten reliability in exactly the places where load is growing.

B Ancillary-Service Obligations in ERCOT

ERCOT is not FERC-jurisdictional, so its ancillary services do not appear in the EQR. We instead use ERCOT's 60-Day SCED Disclosure Reports, which record, for every generation resource and every 15-minute dispatch interval, the megawatts of each ancillary service the resource is responsible for providing. We average these obligations over the intervals of each hour and the hours of each quarter, giving a resource-by-quarter panel for 2014Q2–2025Q3; the panel ends before ERCOT's move to real-time co-optimization in December 2025, which changed how ancillary services are procured. The zone-level design of the main analysis cannot be applied within ERCOT: ERCOT contains only four zones, and none of them crosses the treatment threshold within a window that admits it to a stack. We therefore apply the same estimator at the level of individual resources. A resource is treated in the first quarter in which operating data center capacity within 30 km rises a given amount above its 2014Q1 level; controls are resources that are never treated. The event window, fixed effects, stack weights and inference follow equation (1), with standard errors clustered by three-digit ZIP code because neighboring resources share the same exposure.

Figure IA1 reports estimates with a 100 MW threshold. After activation, fossil resources near data centers carry 1.22 MW more Responsive Reserve ($p = 0.014$) and 0.52 MW more regulation-up ($p = 0.040$), relative to pre-activation means among treated resources of 0.64 MW and 0.45 MW. For all resources, the regulation-up estimate is similar (0.39 MW, $p = 0.046$) and the Responsive Reserve estimate is positive but not significant (0.72 MW, $p = 0.18$). None of these estimates shows a significant pre-activation coefficient. Both services can be provided only by units that are online and running below their maximum output, so the increase is consistent with additional local load keeping nearby fossil units committed and able to carry reserves. These results should be read as suggestive. With the 20 MW threshold used in the main analysis, no ancillary-service estimate is significant at the 5% level, and the estimates change sign across thresholds. Energy output of the same resources is higher both before and after activation, so the design does not identify an effect on output in ERCOT.



(a) All: Responsive Reserve

(b) All: Regulation up

(c) Fossil: Responsive Reserve

(d) Fossil: Regulation up

Figure IA1: **Event-study estimates: ancillary-service obligations in ERCOT (SCED resource × quarter).** Notes: The figure presents event-study estimates for ERCOT generation resources, 2014Q2–2025Q3, from ERCOT's 60-Day SCED Disclosure Reports. The dependent variable is the resource's average obligation (MW) for the service, averaged over the 15-minute intervals of each hour and the hours of each quarter. A resource is treated in the first quarter in which operating data center capacity within 30 km rises at least 100 MW above its 2014Q1 level; controls are resources that are never treated. Rows: all resources; fossil resources (simple-cycle turbines, diesel, combined cycle, coal and steam units). Event time is in quarters, $e \in [-8, 16]$, and $e = -1$ is omitted. Regressions include resource × stack and stack × quarter fixed effects; standard errors are clustered by three-digit ZIP code, and whiskers are 95% confidence intervals.