# Per- and polyfluoroalkyl substances (PFAS) contamination following data center expansion*

Xinming Du† Douglas Almond‡

February 2026

**Abstract**

Data center expansion generates energy and land use concerns, yet its direct pollution impacts remain understudied. Using granular data on data center siting and commissioning, together with high-frequency water pollution and labor data, we show that data centers in the U.S. account for 1.2 percent of current per- and polyfluoroalkyl substances (PFAS) contamination levels. On the labor side, each additional data center increases local employment by 3.1 percent, driven by expansion in service-oriented sectors. However, hours worked and wages remain unchanged, indicating that workers do not avoid work and are not compensated through higher pay, despite rising pollution exposure.

JEL Classifications: Q53, Q51, J23, R11, Q56

Keywords: data centers, persistent chemicals, PFAS, water pollution, employment

*We thank Hannah Farkas and Junho Choi for helping with the Homebase data.

†Department of Economics, National University of Singapore. Email: xdu@nus.edu.sg

‡School of International and Public Affairs & Department of Economics, Columbia University & NBER. Email: da2152@columbia.edu

# 1 Introduction

The rapid expansion of data centers has become a defining feature of the modern digital economy. These facilities underpin cloud computing, artificial intelligence, e-commerce, and a wide array of data-intensive services that drive productivity, innovation, and economic growth. Governments and private firms increasingly view digital infrastructure as a strategic investment, and the scale of data center construction has accelerated sharply in recent years. At the same time, data centers impose growing demands on local resources, including electricity, land, and water. Existing research has focused primarily on their implications for energy systems, grid reliability, and carbon emissions, yet far less is known about their broader environmental externalities, particularly those arising through less visible channels such as chemical pollution.

This paper studies an overlooked environmental consequence of digital infrastructure: contamination from per- and polyfluoroalkyl substances (PFAS) associated with data center expansion. PFAS are highly persistent synthetic chemicals used in industrial processes and consumer products for their heat resistance and chemical stability. These same properties make PFAS difficult to degrade, leading to long-term accumulation in soil and water and widespread exposure through drinking water systems. A growing body of medical and economic evidence links PFAS exposure to adverse health outcomes, reduced human capital, and large economic costs, including declines in property values and increased public spending on monitoring and remediation. Despite increasing regulatory attention worldwide, there is limited empirical evidence on how emerging sources of PFAS linked to modern infrastructure contribute to ongoing contamination.

Data centers represent a plausible but under-recognized source of PFAS pollution. Modern facilities rely on advanced cooling technologies, electrical insulation, and specialized electronic components that frequently incorporate PFAS-based substances for their thermal stability and dielectric properties. Leaks, maintenance failures, waste handling, and upstream production processes associated with these inputs can release PFAS into surrounding soil and groundwater. Because data centers are often located based on cost and infrastructure considerations rather than proximity to final users, local communities may bear environmental and health costs without directly benefiting from the services these facilities provide. Understanding this trade-off is essential for evaluating the true social returns to digital infrastructure investments.

We provide the first systematic evidence linking data center expansion to PFAS contamination in drinking water. Using granular facility-level data on the timing and location of

U.S. data centers combined with high-frequency monitoring data from the EPA's Unregulated Contaminant Monitoring Rule (UCMR-5), we exploit within-zipcode variation in data center openings and expansions. Our empirical strategy employs panel regressions with rich fixed effects and event-study designs to isolate causal effects and assess dynamic responses around the timing of data center entry.

Our results show that data center activity leads to statistically and economically significant increases in PFAS contamination. The entry of one additional data center raises PFAS concentrations by 10 percent relative to the sample mean and increases the probability of PFAS detection by 9 percent. These effects are concentrated among long-chain PFAS compounds such as PFOS and PFOA, which are more persistent and more strongly linked to adverse health outcomes. Effects are also larger for facilities with on-site power generation, indicating that operational scale and design features play an important role in determining environmental impacts. Event-study estimates reveal no evidence of differential pre-trends and show that PFAS levels rise persistently following data center opening, consistent with ongoing release and accumulation rather than temporary construction-related shocks.

We further examine whether data center expansion affects other dimensions of local welfare. Consistent with chemically driven pollution pathways, we find that data center entry increases conventional measures of water pollution such as chemical oxygen demand, with weaker effects on biodegradable pollutants and suspended solids. On the economic side, data centers generate large local employment gains, primarily in service-oriented sectors such as transportation, food services, and facility support. However, these gains are not accompanied by increases in wages or hours worked, suggesting job creation without meaningful compensation premia or skill upgrading.

Our paper contributes to the literature on the impacts of data center expansion and to the emerging economics of PFAS. In the data center literature, prior work has focused on location choice and supply chain design, modeling optimal siting and capacity allocation under latency and cost trade-offs (Faiz and Noor-E-Alam, 2019; Liang et al., 2021), as well as on how data-center expansion and operational flexibility affect grid operation, investment, and emissions outcomes (Bogmans et al., 2025; Knittel et al., 2025; Shawhan et al., 2025). A related body of work documents the sector's rapidly rising electricity demand and cooling intensity, highlighting environmental concerns but largely relying on modeling exercises or descriptive evidence without direct empirical measurement of localized pollution impacts (International Energy Agency, 2024a; Masanet et al., 2020; Wang and Lv, 2024; Bonner et al., 2024). We extend this literature by providing causal evidence that data center expansion generates measurable

chemical pollution, identifying a previously unquantified environmental externality of digital infrastructure.

We are also among the first to study PFAS using large-scale, high-frequency data in an economic context. While a growing scientific and public health literature documents the persistence and health risks of these "forever chemicals" (e.g. Hu et al., 2016; Andrews and Naidenko, 2020; Fenton et al., 2021), economic studies remain limited and focus on specific point sources and information shocks. Existing work shows that PFAS exposure lowers human capital and earnings using quasi-experimental variation from firefighting foam at military installations and that public revelations of contamination depress housing prices (Jacqz et al., 2024; Marcus and Mueller, 2024). Baluja et al. (2025) exploit variation in drinking water flows in New Hampshire and find that downstream PFAS exposure increases infant mortality, preterm births, and low birthweight. We complement and extend this literature by focusing on PFAS presence in surface water rather than other environmental pathways. We also contribute by linking nationwide PFAS monitoring data to the rapid expansion of data centers. Our findings reveal a new, technologically driven source of exposure that is relevant for contemporary debates on sustainable AI, infrastructure investment, and environmental regulation.

From the policy perspective, our findings highlight a fundamental trade-off in the expansion of digital infrastructure. While data centers contribute to economic growth and employment, they also generate persistent environmental externalities that are not currently priced or fully regulated. By documenting a causal link between data center expansion and PFAS contamination, this paper provides new evidence on the hidden environmental costs of the digital economy and underscores the importance of integrating chemical pollution into infrastructure planning and environmental policy. Incorporating these externalities is essential for designing regulatory frameworks that ensure the transition to a digitally driven economy does not come at the expense of long-run environmental quality and public health.

## 2 Background

### 2.1 Expanding data centers

Data centers are physical facilities that house servers, networking equipment, and storage systems used to process, store, and transmit digital information. While data centers have existed for decades, their role within the digital economy has shifted markedly in recent years. Earlier generations primarily supported enterprise computing and basic data storage.

In contrast, modern data centers now serve as the core infrastructure for cloud computing, large-scale data processing, and especially the training and operation of artificial intelligence systems. Their primary purpose is therefore no longer passive storage, but the continuous execution of computation-intensive digital services.

Although data centers are not new, their expansion has accelerated sharply in recent years. Global electricity consumption by data centers reached 415 TWh in 2024, accounting for around 1-1.5% of total global electricity demand, and is projected to more than double by 2030 under current trends (International Energy Agency, 2024b). Cooling systems alone account for 40% of total energy consumption within data centers, highlighting their great operational intensity (Masanet et al., 2020).

Importantly, data centers do not need to be located close to final users. While latency considerations influence network design, modern distributed systems and traffic-routing architectures allow computation and storage to occur far from end users without large performance loss. Consequently, location decisions are primarily shaped by cost and infrastructure factors, including access to inexpensive electricity, suitable land, and reliable grid capacity, rather than proximity to consumer markets (Faiz and Noor-E-Alam, 2019; Liang et al., 2021).

This has reshaped the spatial pattern of data centers. While earlier facilities were often concentrated near corporate or urban business centers, recent construction has increasingly shifted toward regions offering low-cost power and large tracts of developable land. In the U.S., major clusters have emerged in locations such as Northern Virginia, where installed capacity exceeds 2,500 MW, as well as in Texas, Arizona, and parts of the Midwest. At the national level, data centers are widely distributed, but construction activity has accelerated notably after 2020, reflecting growing demand induced by AI and cloud computing.

The rapid expansion and geographic concentration of data centers have raised growing concerns regarding grid congestion, fossil-fuel lock-in, and other environmental externalities (Bogmans et al., 2025; Knittel et al., 2025; Shawhan et al., 2025). While most existing studies focus on energy demand and grid implications, the proliferation of data centers also raises important questions about less visible environmental pathways.

### 2.2 Persistent PFAS

Per- and polyfluoroalkyl substances are a large class of synthetic chemicals that have been used since the 1940s in non-stick coatings, waterproofing agents, stain-resistant textiles, and fire-fighting foams. Their exceptionally strong carbon-fluorine bonds make them highly resistant

to thermal, chemical, and biological degradation. As a result, PFAS persist in the environment for decades, bioaccumulate in humans and wildlife, and are now detected globally in soil, air, surface water, groundwater, and human blood serum (Hu et al., 2016; Andrews and Naidenko, 2020; Austin and Mueller, 2025).

Drinking water is a dominant exposure pathway for the general population. PFAS released from manufacturing facilities, fire-training sites, metal-plating operations, textiles, and wastewater treatment plants can migrate into surface water and aquifers, where they are difficult and costly to remove (Hu et al., 2016; Calafat et al., 2019; Agency for Toxic Substances and Disease Registry, 2020; Andrews and Naidenko, 2020). Once in the hydrological system, PFAS are transported over long distances and can contaminate public water supplies far from the original source.

A large toxicology and epidemiology literature documents adverse health effects induced by PFAS exposure. Documented outcomes include elevated risks of kidney and testicular cancers, immunotoxicity and reduced vaccine response, thyroid dysfunction, altered lipid metabolism, cardiovascular disease, and adverse birth and developmental outcomes (Barry et al., 2013; Rappazzo et al., 2017; Fenton et al., 2021). Given their persistence and bioaccumulation, even low-level chronic exposure may translate into large long-run health burdens.

The economics literature on PFAS is more recent but growing. Jacqz et al. (2024) exploit PFAS-based firefighting foam use at U.S. military sites as a natural experiment and find that in utero exposure lowers birthweight and leads to worse educational and earnings outcomes later in life. Marcus and Mueller (2024) study how information about unregulated PFAS contamination affects housing markets and document declines in property values following public revelations of contamination. These studies highlight that PFAS pollution generates sizable and persistent economic costs in addition to health damages.

Reflecting these concerns, regulatory responses have accelerated in recent years. In the U.S., the Environmental Protection Agency (EPA) finalized the first enforceable national PFAS drinking water standards in April 2024 and continues nationwide monitoring under Unregulated Contaminant Monitoring Rule (UCMR-5).[1] In the European Union, the 2021 Drinking Water Directive sets a limit for total PFAS of 0.5 $\mu$g/L, and a 2023 REACH restriction proposal seeks to phase out most non-essential PFAS uses by 2030. Australia's PFAS National Environmental Management Plan, updated in 2023, provides guidance for investigation, cleanup, and reuse of contaminated sites, and from July 2025 the manufacture, import, and use of PFOS, PFOA, and PFHxS will be banned. In Canada, Health Canada proposed in 2023 a nationwide drinking

[1] Details on the U.S. regulatory timeline are summarized in Austin and Mueller (2025).

water objective of 30 ng/L for the sum of 25 PFAS compounds. These policies indicate that PFAS are now treated as a global regulatory priority, but they also leave open the question of how emerging sources such as expanding digital infrastructure contribute to ongoing contamination.

### 2.3 Why data centers may generate PFAS

Although data centers are not traditionally viewed as sources of chemical pollution, several technical features of modern facilities create plausible pathways for PFAS release. Similar to the historical use of chlorofluorocarbons (CFCs) in refrigeration systems, PFAS-based compounds are widely used in data center cooling technologies, electrical insulation, and electronic components due to their high thermal stability, chemical resistance, and dielectric properties.

One key channel operates through cooling systems. As data centers scale up and adopt increasingly dense server configurations, efficient thermal management becomes critical. Two-phase liquid-cooling systems, which are now increasingly deployed in high-performance computing and AI-oriented facilities, rely on PFAS-based refrigerants and heat-transfer fluids (Bonner et al., 2024). Leaks, maintenance failures, or improper handling of these fluids can result in direct releases into surrounding soil and groundwater, particularly in facilities without fully sealed containment systems.

A second channel arises from the lifecycle of electronic components and cabling. PFAS are used in wire coatings, circuit boards, and semiconductor manufacturing processes to improve durability and performance. Disposal or recycling of electronic waste from data centers can therefore introduce PFAS into landfills or incineration streams. These processes generate persistent byproducts that can migrate into water systems through leachate or atmospheric deposition.

Third, upstream production processes linked to data center supply chains may also contribute to local PFAS contamination. Semiconductor fabrication plants and cooling-fluid manufacturing facilities are well-documented industrial users of PFAS-based substances. As demand for data center infrastructure expands, PFAS emissions from these upstream producers may increase in regions hosting data center clusters.

To sum up, these mechanisms suggest that data centers may represent an emerging and under-recognized source of PFAS pollution. Unlike traditional point sources such as firefighting training sites or chemical manufacturing plants, data centers introduce PFAS through a

combination of operational, waste-management, and supply-chain channels.

# 3 Data

## 3.1 Data centers

Our data on data centers are obtained from Aterio, which provides facility-level information on data centers across the U.S. The dataset is cross-sectional in the sense that each observation represents a unique data center campus, but it contains dated fields, including construction, activation, and expansion dates, that allow us to construct a zipcode-by-time panel of data center activity. For each facility, we observe company name, ownership status (public or private), and a website link. Facility characteristics include building and campus names, geographic coordinates, construction and activation dates, total floor space, number of buildings on the campus, installed power capacity measured in megawatts (MW), and Power Usage Effectiveness (PUE), which captures the ratio of total facility energy consumption to IT equipment energy use.

The dataset also records major events associated with each data center, such as land acquisitions, expansions, and the signing of Power Purchase Agreements (PPAs), particularly for large providers such as Microsoft and Google. These event records allow us to track the timing of center openings and capacity changes over time.

In total, the dataset contains 5,145 data centers at any stage of development, of which 1,907 are currently active. Data centers are widely distributed across the U.S., as shown in Figure 1. While facilities are geographically dispersed, construction has increasingly gravitated toward regions offering low-cost electricity and large tracts of developable land.

Although data centers are not a new phenomenon, construction activity has accelerated markedly in recent years. Figure S1 shows that the number of new data centers increased sharply after 2020, consistent with the rapid expansion of cloud computing and AI-related demand.

## 3.2 PFAS

Our PFAS data come from the U.S. EPA's Unregulated Contaminant Monitoring Rule (UCMR-5), which mandates nationwide monitoring of 29 PFAS substances and lithium in public water

systems (PWSs). The program covers 7,350 PWSs and 19,500 sampling locations, representing 83% of all U.S. public water systems.

The monitoring period spans from January 3, 2023 to June 25, 2025, providing high-frequency observations over a relatively short time window. Each raw record contains a PWS identifier, 5-digit zipcode, sample collection date, contaminant identifier (one of the 29 PFAS compounds), the sampled concentration when detected, and the minimum reporting level (MRL).

We collapse the raw data to the PWS-zipcode-pollutant-day level and construct three alternative measures of PFAS exposure to address left-censoring arising from non-detection. The first measure replaces missing values with zeros, treating non-detects as true zeros. The second replaces missing values with half of the MRL, reflecting a common midpoint imputation approach. The third is a detection indicator that takes the value one if PFAS are detected and zero otherwise. These alternative constructions allow us to assess the robustness of our results to different assumptions about non-detect observations, which are pervasive in PFAS monitoring data. In our sample, the mean detection rate is 2.3%, implying that PFAS presence in drinking water is far from rare, even within the limited monitoring horizon. These PFAS measures are then linked spatially and temporally to data center activity at the 5-digit zipcode level.

### 3.3 Water pollution

To examine broader water quality impacts induced by data center development, we compile surface water pollution data from the U.S. EPA's STORET database and the U.S. Geological Survey's National Water Information System (NWIS). We follow the aggregation procedure in Keiser and Shapiro (2019) to harmonize measurements across monitoring sites and construct consistent zipcode-level pollution indicators.

The raw data are recorded at the monitor-day level for each pollutant. Using the geographic coordinates of monitoring stations, we spatially aggregate observations to the 5-digit zipcode-day level and then to the zipcode-month level for the main analysis. Our primary water quality outcomes are chemical oxygen demand (COD), biochemical oxygen demand (BOD), and suspended solids (SS), which capture organic pollution, biological contamination, and particulate matter in surface water, respectively.

These data span a long time window from January 2001 to June 2025, allowing us to examine both short- and medium-run changes in water pollution following data center openings. The use of multiple pollutants enables us to assess whether data center development results in a

general deterioration in water quality or whether effects are concentrated in specific dimensions of pollution. We link these zipcode-level water pollution measures to data center activity over time, constructing a panel that allows us to estimate changes in local water quality as data centers enter and expand.

### 3.4 Labor outcomes

Our labor market data come from Homebase, which provide high-frequency information on worker schedules and compensation across a wide range of service-sector establishments in the U.S. For each worker-day observation, the data report total wages earned, hours worked, and the average hourly wage, along with anonymized worker identifiers, worker level (employee or manager), firm identifiers, county, zipcode, and NAICS industry classification.

We aggregate the raw worker-day observations to the industry-zipcode-day level, covering the period from June 28, 2020 to October 28, 2025. Our primary labor market outcomes include the total number of workers, working hours per worker, total earnings per worker, and hourly wages. These measures allow us to examine both extensive-margin responses, such as employment changes, and intensive-margin responses, such as adjustments in hours and compensation.

The richness of the Homebase data enables us to explore labor market effects across worker levels and industry types, providing insight into whether data center development affects local employment patterns through increased labor demand, changes in business activity, or shifts in local economic conditions. We link these labor outcomes spatially and temporally to data center openings and expansions at the zipcode level.

## 4 Empirical strategy

We examine whether data center development leads to changes in PFAS contamination by exploiting spatial and temporal variation in the opening and expansion of data centers across U.S. zipcodes. Our baseline specification is estimated at the PWS-zipcode-pollutant-day level and is given by:

$$PFAS_{ijpt} = \beta \cdot Center_{jt} + \alpha_p + \gamma_i + \delta_j + \lambda_t + \varepsilon_{ijpt} \tag{1}$$

where $PFAS_{ijpt}$ measures PFAS contamination for public water system $i$, located in zipcode $j$, for pollutant $p$, on day $t$. We use three PFAS outcome measures: concentration with missing

values filled with zero, concentration with missing values replaced by half of the minimum reporting level, and a detection indicator.

On the right-hand side, the variable $Center_{jt}$ denotes the number of active data centers in zipcode $j$ at time $t$. In alternative specifications, we replace this count variable with total installed data center capacity measured in megawatts (MW) to capture intensity effects. The specification includes pollutant fixed effects $\alpha_p$ to account for systematic differences across PFAS compounds, public water system fixed effects $\gamma_i$ to absorb time-invariant characteristics of monitored systems, and zipcode fixed effects $\delta_j$ to control for stable geographic differences in baseline contamination risk and infrastructure. Time fixed effects $\lambda_t$ are flexibly specified using combinations of year, month, day-of-week, or full date indicators to capture common shocks, seasonality, and changes in monitoring regimes. Standard errors are clustered at the PWS-zipcode level to allow for serial correlation in outcomes within the same water system and geographic area over time.

Identification relies on variation in data center activity within a zipcode over time, under the assumption that, conditional on the included fixed effects, changes in data center presence are orthogonal to unobserved time-varying determinants of PFAS concentrations. To test this assumption, we estimate event-study specifications centered on the timing of data center openings. These models allow us to assess whether PFAS levels exhibit systematic changes prior to center entry and to trace the evolution of contamination following activation.

# 5 Results on PFAS

## 5.1 Main results

Table 1 presents our main estimates on the relationship between data center activity and PFAS contamination at the PWS-zipcode-pollutant-day level. Panels A-C measure exposure using the number of active data centers, while Panels D-F use total installed data center capacity (MW) to capture the intensive margin of data center operations.

In Panel A, where PFAS concentration is constructed by filling missing values with zero, the preferred specification implies that an increase of 1,000 data centers within a zipcode raises PFAS concentration by 0.018 $\mu$g/L. Scaled to more interpretable units, the addition of one data center increases PFAS concentration by approximately $10.6\,\%$ relative to the sample mean and by $0.85\,\%$ relative to the sample standard deviation. These estimates remain highly stable as we sequentially introduce increasingly stringent fixed effects, including PWS fixed effects,

zipcode fixed effects, PWS-zipcode fixed effects, and full date fixed effects, suggesting that the relationship is not driven by unobserved time-invariant characteristics of water systems or locations, nor by common temporal shocks.

Panel B shows that the magnitude and statistical significance of the estimates are highly similar when missing PFAS observations are replaced with half of the minimum reporting level (MRL), indicating that the results are not sensitive to the treatment of non-detects or to censoring at low concentration levels. This consistency strengthens confidence that the observed relationship reflects genuine differences in contamination levels rather than mechanical features of the measurement process.

Panel C examines the extensive margin using a PFAS detection indicator. The results show that data center presence significantly increases the probability that PFAS are detected in nearby water systems. The addition of one data center raises the detection probability by 0.0021, corresponding to 9.26% of the sample mean and by 1.55% relative to the standard deviation. Given a mean detection rate of 0.023, this implies a large relative increase in the likelihood that PFAS exceeds reporting thresholds in areas with data center activity.

Panels D-F repeat the analysis using total installed data center capacity (MW) instead of the number of centers, allowing us to assess whether PFAS contamination scales with operational intensity. In Panel D, installed data center capacity is measured in terawatts (TW). A one-terawatt increase in capacity raises PFAS concentration by $0.357\ \mu\text{g/L}$ when missing values are filled with zero. Scaled to a more interpretable magnitude, a $1{,}000\ \text{MW}$ (i.e., $0.001\ \text{TW}$) increase in capacity implies a $0.00036\ \mu\text{g/L}$ increase in PFAS concentration, corresponding to a $0.21\%$ increase relative to the sample mean and a $0.17\%$ increase relative to one standard deviation. Interpreted at smaller scales, a 100 MW increase in capacity — comparable to the size of a large hyperscale facility — implies a non-trivial rise in PFAS concentration, underscoring the relevance of capacity expansion even within already-developed data center regions.

Panel E shows that when concentrations are constructed using half of the MRL for missing values, a 1,000 MW increase in capacity raises PFAS concentration by 0.020 $\mu$g/L, equivalent to a 9.4% increase relative to the mean and a 0.75% increase relative to the standard deviation. The near-identical magnitude across Panels D and E further reinforces the robustness of the capacity results and confirms that the intensive-margin relationship is not an artifact of how non-detects are handled.

Panel F focuses on the detection margin and shows that greater installed capacity also significantly increases the likelihood that PFAS are detected. A 1,000 MW increase in capacity

raises detection probability by 0.0021, corresponding to a 9.13% increase relative to the mean detection rate and a 1.48% increase relative to the standard deviation. This indicates that larger and more power-intensive data centers not only increase average PFAS concentrations, but also shift water systems across regulatory detection thresholds more frequently.

Figure 2 provides complementary dynamic evidence by tracing PFAS concentration around the timing of data center openings. The event-study specification uses PFAS concentrations constructed by replacing missing values with half of the MRL and includes contaminant, PWS, zipcode, year, month, and day-of-week fixed effects, with standard errors clustered at the PWS-zipcode level. The pre-treatment coefficients are small, flat, and statistically indistinguishable from zero, indicating no evidence of differential pre-trends prior to data center entry. This suggests that treated and untreated zipcodes followed similar trends in PFAS contamination before centers opened, supporting the identifying assumption underlying the main regressions.

Following data center entry, PFAS concentrations increase steadily and persistently. The effect emerges shortly after opening and remains positive throughout the post-treatment window, with no sign of mean reversion. This dynamic pattern is consistent with a mechanism in which PFAS are released continuously through data center operations and gradually accumulates in local groundwater or surface water systems, rather than reflecting a one-time construction shock.

### 5.2 Heterogeneity

We examine heterogeneity in the PFAS response to data center activity across individual PFAS chemicals. Figure S2 presents estimates separately for each PFAS compound. The effects are highly uneven across chemicals. The largest and most precisely estimated responses are observed for long-chain PFAS compounds, including PFOS, PFOA, PFHxS, and PFNA. These compounds exhibit coefficients close to or exceeding 0.10-0.15 $\mu$g/L per 1,000 data centers, larger than the near-zero or statistically insignificant effects observed for short-chain compounds such as PFBA, PFBS, and PFPeA. For example, the point estimate for PFOS exceeds 0.15 $\mu$g/L, while several short-chain chemicals cluster tightly around zero and are statistically indistinguishable from no effect. This pattern indicates that the aggregate PFAS increase documented in the main results is driven disproportionately by a subset of high-risk, long-chain compounds rather than by a uniform shift across all PFAS species. The chemical-specific pattern is consistent with differences in industrial use and persistence, with long-chain PFAS more likely to accumulate in environmental media and human water systems.

We next explore heterogeneity by on-site power generation, motivated by concerns that self-generation capacity may intensify environmental impacts. Table S1 reports regressions interacting data center exposure with an indicator for facilities operating on-site generators. The results reveal large and economically meaningful differences. When exposure is measured by the number of data centers, a 1,000-center increase with on-site generators raises PFAS concentration by 0.028-0.029 $\mu$g/L, while the corresponding estimate for centers without generators is close to zero and statistically insignificant. This implies that the PFAS response for data centers with on-site power is roughly double that of the baseline average effect and several times larger than that for centers without on-site generation.

A similar pattern holds when exposure is measured using installed capacity. A 1,000 MW increase in capacity with on-site generation leads to a 0.97 $\mu$g/L increase in PFAS concentration, whereas capacity expansions among centers without generators yield coefficients between 0.01 and 0.13 $\mu$g/L and are not statistically distinguishable from zero. These differences imply that the intensive PFAS response is concentrated among large-scale, power-intensive data centers that operate self-generation systems.

### 5.3 Incorporating river flow networks

PFAS contaminants can travel through hydrological systems, implying that public water supplies may be exposed to pollution originating from data centers located upstream. To capture this channel, we extend our empirical strategy by incorporating information on river-flow direction. This practice also helps address potential concerns that data center siting or operational timing may be correlated with other events occurring within the same zipcode, but are far less likely to be correlated with conditions in upstream zipcodes connected only through river networks.

We obtain hydrologic connectivity information from the National Hydrography Dataset Plus (NHDPlus), a nationwide hydrography dataset jointly produced by the EPA and the USGS. NHDPlus combines detailed river and stream networks with catchment boundaries, digital elevation models, and flow-direction attributes that enable tracing of surface water movement through interconnected reaches. Each river segment is assigned a unique identifier and mapped to its upstream and downstream neighbors, allowing us to construct directed river-flow paths across the national water system. Using these connectivity attributes, we determine for each ordered pair of zipcodes $(k, j)$ whether water flows from zipcode $k$ to zipcode $j$ and count the number of distinct flow paths contributing to potential contaminant

transport between them. We then construct a normalized exposure measure that captures the weighted average number of upstream data centers connected to zipcode $j$:

$$\#UpstreamCenter_{jt} = \frac{\sum_k \#UpstreamRivers_{kj} \cdot \#Center_{kt}}{\sum_k \#UpstreamRivers_{kj}}$$

We then estimate the following specification:

$$PFAS_{ijpt} = \beta \cdot \#UpstreamCenter_{jt} + \alpha_p + \gamma_i + \delta_{jt} + e_{ijpt}, \tag{2}$$

where the sample is the same as in Equation (1) at the PWS-zipcode-pollutant-day level from January 2023 to June 2025. In robustness exercises, we additionally include data centers located in the same zipcode as a horse-race term, to account for the potential concern that PFAS levels may be positively correlated across neighboring locations. We also replace upstream exposure with downstream exposure as a directional placebo test.

Results are reported in Table 2. Panel A measures exposure using the number of active upstream data centers. An increase of 1,000 upstream data centers raises PFAS concentrations by 0.000087 $\mu$g/L when missing values are imputed as zero, corresponding to 50.6 percent of the sample mean and 4.1 percent of the sample standard deviation. Using half of the minimum reporting level yields similar magnitudes, corresponding to approximately 48 percent of the mean. For PFAS detection, the same exposure increases the detection probability by 0.00513, corresponding to a 0.51 percentage point increase, which equals 22.4 percent of the mean and 3.7 percent of the standard deviation.

Panel B replaces the count-based exposure with installed upstream data center capacity. The estimated coefficients are comparable in magnitude and precision to those in Panel A, indicating that PFAS contamination scales with upstream operational intensity rather than merely the number of facilities.

Panels C and D additionally control for the number or capacity of data centers located in the same zipcode. Upstream coefficients remain positive and statistically significant, declining only modestly in magnitude. Same-zipcode coefficients are consistent with the estimates reported in Section 5.1, indicating that PFAS contamination arises both from local emissions and through upstream hydrological transport. This pattern highlights that environmental externalities propagate across administrative boundaries, with downstream communities affected by facilities located upstream.

Panels E and F implement a placebo test using downstream data centers, which by construction should not affect upstream PFAS concentrations. The estimated coefficients are small and statistically indistinguishable from zero across all outcome definitions. This falsification exercise supports a hydrological interpretation of the upstream effects rather than correlated siting decisions or broader spatial trends.

Our results document a clear river-flow pollution spillover channel. PFAS released by data centers in one location can elevate contamination levels at downstream water systems that do not host data centers themselves. These findings imply that the environmental externalities of data center expansion extend beyond local administrative boundaries, and that downstream communities may bear pollution risks without associated economic benefits.

# 6 Results on other environmental outcomes

## 6.1 Correlation between PFAS and water pollution

In this section, we examine the raw association between PFAS contamination and other water pollution outcomes. We first analyze the raw relationship between PFAS contamination and conventional measures of water quality, focusing on chemical oxygen demand (COD), biochemical oxygen demand (BOD), and suspended solids (SS). Figure S3 presents binned scatter plots for the period January 2023 to June 2025, without including fixed effects or additional controls.

Across all three PFAS constructions, we observe a clear positive correlation between PFAS and COD. This positive relationship remains stable when PFAS are measured using concentrations with missing values filled with zero, with half of the minimum reporting level (MRL), and when using the detection indicator. A similar pattern emerges for BOD. For suspended solids, the association with PFAS are also positive, though less precise and smaller in magnitude. This likely reflects the fact that suspended solids capture a broader category of particulate pollution, some of which may not be directly linked to PFAS transport pathways. These patterns suggest that PFAS contamination does not appear spatially isolated but instead co-moves with other indicators of deteriorating water quality.

## 6.2 Effects on water pollution

We next examine whether data center expansion affects conventional measures of water quality. The results are reported in Tables S2, which exploits panel variation at the zipcode-month

level from January 2001 to June 2025. For COD, additional data centers result in higher concentrations in both level and proportional specifications. An additional data center raises COD by 1.15 mg/L, implying a 7 percent increase relative to the sample mean. When exposure is measured using installed capacity, a 1 gigawatt increase in capacity raises COD by 40 mg/L, indicating a large increase in organic pollutant load. These effects remain stable when switching between models with separate year and month fixed effects and those with combined year-month fixed effects, suggesting that the results are not driven by differential macro trends or seasonality.

A similar but weaker pattern emerges for BOD. Data center activity leads to higher biochemical oxygen demand, with an additional data center increasing BOD by around 0.14 mg/L, although this estimate is less precise. Capacity-based estimates suggest a modest response: a 1 gigawatt increase in installed capacity corresponds to a 10 percent increase in BOD levels. The weaker response of BOD relative to COD likely reflects the nature of pollutants generated by data centers. COD captures both biodegradable and non-biodegradable organic matter, including chemically persistent compounds such as PFAS and industrial coolant byproducts, whereas BOD specifically measures oxygen consumption driven by microbial decomposition of biologically degradable material. Since data centers primarily release chemically stable and low-biodegradability substances rather than biologically reactive organic waste, increases in COD are more pronounced than increases in BOD.

Suspended solids also show a weaker response to data center expansion. While point estimates are generally positive, they are not consistently statistically significant across specifications, indicating no strong or systematic change in particulate concentrations. This muted response is consistent with the operational characteristics of data centers, which generate relatively limited sediment runoff or physical particulate pollution compared to construction-heavy or extractive industries. Once operational, data centers are largely enclosed facilities, and their primary environmental impacts operate through liquid discharges and chemical effluents rather than erosion, soil disturbance, or particulate waste streams.

Figure S4 provides dynamic evidence using an event-study design for COD. The pre-treatment coefficients are close to zero and statistically insignificant, suggesting no differential trends prior to data center opening. Following entry, COD concentrations increase and remain persistently elevated over time, indicating that the effect reflects a sustained deterioration in water quality rather than a short-lived construction-related shock.

### 6.3 Discussion on health risks

Our findings provide robust evidence that data center operations increase surface water pollution, especially through elevated COD and PFAS concentrations. These patterns suggest potential risks to human health. Quantifying such risks is challenging because actual exposure depends on multiple factors, including piped water coverage, treatment quality and cost, bottled water consumption, and recreational contact with contaminated waterways. Nevertheless, the existing literature provides estimates that allow us to benchmark our findings.

The closest setting is Flynn and Marcus (2025), who study the Clean Water Act in the U.S. and show that treatment-technology upgrades decreased dissolved oxygen deficit by 1.6 pp downstream, which in turn increased average birth weight by 8 g and reduced the incidence of low birth weight (LBW) by 1-4%. Other studies on surface water and health provide reduced-form evidence that upstream activities, agricultural runoff, or disease-related contamination increase disability (Lai, 2017), diarrhea (Garg et al., 2018), short-term mortality (Jones, 2019), and long-term mortality (Fletcher and Noghanibehambari, 2024), but they do not offer comparable first-stage estimates of how water quality itself changes.

Our estimates indicate that each additional data center located in the same zipcode increases COD concentrations by $1.10\ \mathrm{mg/L}$ at the zipcode-month level ($4.6\%$ of the mean) and increases PFAS concentrations by $0.00002\ \mu\mathrm{g/L}$ at the zipcode-day level ($10.6\%$ of the mean). Assuming the same elasticity of birth outcomes to surface water quality as in Flynn and Marcus (2025)—which is a very strong assumption since dissolved oxygen deficit and COD/PFAS capture different pollutants and exposure differs across settings—these pollution increases imply birth weight losses of 5-20 g and a 1-9% increase in LBW, as well as potential increases in disability, diarrheal illness, and infant, short-, and long-term mortality.

## 7 Labor market outcomes

### 7.1 Effects on employment, hours worked, and wages

Table S3 reports the effects of data center openings on local labor outcomes at the industry-zipcode-day level. The most consistent and robust effect is on employment. Each additional data center increases the number of workers by 3.1-3.14%, statistically significant at the 1% level. Given a mean log worker count of 4.44, this represents a meaningful expansion of local employment. Using total data center capacity yields a similarly strong result: an additional

gigawatt of capacity raises employment by 67%, highlighting the high labor footprint of large-scale facilities.

In contrast, we find no statistically significant effects on hours worked per worker. The coefficient on the number of data centers is 0.0019 with a standard error of 0.0014, indicating that the employment response operates through the extensive margin — more workers being hired — rather than through longer working hours.

We also observe no significant effects on total earnings per worker or hourly wages. The estimated impact on hourly wages is close to zero (0.0092-0.0093) and statistically insignificant, and the capacity-based specifications similarly show no robust wage response. This suggests that data center expansion generates job creation without materially altering local wage structures or labor compensation.

To sum up, the identified pattern that rising employment without changes in hours or wages is consistent with a labor demand shock concentrated in routine and support services such as security, maintenance, cleaning, food services, and transportation. These sectors require additional staffing but do not necessarily increase the productivity or bargaining power of existing workers. Firms therefore respond by expanding headcount rather than intensifying labor utilization.

Moreover, these positions are likely filled from an elastic local labor supply, often involving hourly or part-time workers, which dampens upward pressure on wages. The absence of wage effects further indicates that data center-driven job growth does not reflect major skill upgrading or competition for scarce specialized labor, but instead expands employment in relatively standard service occupations. To sum up, our findings suggest that data center expansion appears to function as a job-creation shock with limited spillovers into labor intensity or compensation, generating broad-based employment growth without corresponding increases in hours or pay.

### 7.2 Heterogeneity

Table S4 reports heterogeneous effects by worker level. The increase in employment following data center openings is concentrated among employees. An additional data center raises employee headcount by 2.89 percent and this effect is statistically significant at the one percent level. In contrast, the estimated effects for managers and general managers are small and statistically insignificant. There is no evidence of differential responses across worker levels for working hours, total earnings per worker, or hourly wages. This indicates that data center

expansion primarily increases staffing among lower-tier workers rather than altering labor intensity or compensation.

Table S5 presents heterogeneity across industries. Employment increases are concentrated in food and drink, leisure and entertainment, and transportation. The effects in other sectors including beauty and personal care, education and charities, healthcare and fitness, home repair, professional services, and retail are small and statistically insignificant. Even in industries with positive employment responses, we do not observe corresponding changes in hours worked, earnings, or wages. Overall, these results suggest that data center-driven job creation is concentrated in service-oriented sectors that support local operations rather than reflecting broad changes in job quality or pay.

## 8 Discussion and conclusion

This paper shows that the expansion of data centers generates clear and measurable environmental externalities. Our main estimates indicate that the entry of one additional data center increases PFAS concentration by 0.000022 $\mu$g/L, corresponding to a 10.6% increase relative to the sample mean. Using detection as the outcome, one additional data center raises the probability of PFAS detection by 0.00021, implying a 9.3% increase relative to the mean detection rate. These impacts are concentrated among long-chain PFAS compounds such as PFOS and PFOA and are especially larger for facilities equipped with on-site power generation, highlighting the importance of both scale and operational design in shaping environmental consequences.

To benchmark economic significance, we apply these estimates to the observed distribution of data centers. The average zipcode hosts 0.095 data centers in our sample. This implies that existing data centers raise PFAS concentration by $2.1 \times 10^{-6}$ $\mu$g/L, corresponding to 1.2% of the observed mean concentration. On the extensive margin, the implied increase in detection probability is 0.00019, or 0.9% of the mean detection rate. The results indicate that current data center operations already contribute to measurable increases in long-run PFAS exposure, and continued expansion is likely to amplify contamination risks in the absence of stricter containment and monitoring policies.

## References


**Agency for Toxic Substances and Disease Registry**, "Toxicological Profile for Perfluoroalkyls," Technical Report, U.S. Department of Health and Human Services 2020.

**Andrews, David Q and Olga V Naidenko**, "Population-wide exposure to per-and polyfluoroalkyl substances from drinking water in the United States," *Environmental Science & Technology Letters*, 2020, *7* (12), 931–936.

**Austin, Wes and Rosie Mueller**, "Forever Chemicals: Challenges and Opportunities for Researchers," *Review of Environmental Economics and Policy*, 2025, *19* (1), 118–130.

**Baluja, Robert, Bo Guo, Wesley Howden, Ashley Langer, and Derek Lemoine**, "PFAS-contaminated drinking water harms infants," *Proceedings of the National Academy of Sciences*, 2025, *122* (50), e2509801122.

**Barry, Vaughn, Andrea Winquist, and Kyle Steenland**, "Perfluorooctanoic acid (PFOA) exposures and incident cancers among adults living near a chemical plant," *Environmental health perspectives*, 2013, *121* (11-12), 1313–1318.

**Bogmans, Christian, Patricia Gomez-Gonzalez, Ganchimeg Ganpurev, Giovanni Melina, Andrea Pescatori, and Sneha Thube**, "Power hungry: How ai will drive energy demand," Technical Report 4 2025.

**Bonner, Richard, Bill Grieco, and Liz Cruz**, "Understanding PFAS Concerns for Two-Phase Cooling of Data Centers," Technical Report, Accelsius White Paper 2024.

**Calafat, Antonia M, Kayoko Kato, Kendra Hubbard, Tao Jia, Julianne Cook Botelho, and Lee-Yang Wong**, "Legacy and alternative per-and polyfluoroalkyl substances in the US general population: Paired serum-urine data from the 2013–2014 National Health and Nutrition Examination Survey," *Environment international*, 2019, *131*, 105048.

**Faiz, Tasnim Ibn and Md Noor-E-Alam**, "Data center supply chain configuration design: a two-stage decision approach," *Socio-Economic Planning Sciences*, 2019, *66*, 119–135.

**Fenton, Suzanne E, Alan Ducatman, Alan Boobis, Jamie C DeWitt, Christopher Lau, Carla Ng, James S Smith, and Stephen M Roberts**, "Per-and polyfluoroalkyl substance toxicity and human health review: Current state of knowledge and strategies for informing future research," *Environmental toxicology and chemistry*, 2021, *40* (3), 606–630.

**Fletcher, Jason and Hamid Noghanibehambari**, "The siren song of cicadas: Early-life pesticide exposure and later-life male mortality," *Journal of Environmental Economics and Management*, 2024, *123*, 102903.

**Flynn, Patrick and Michelle Marcus**, "A watershed moment: the Clean Water Act and birth weight," *Journal of Human Resources*, 2025, *60* (6), 1915–1936.

**Garg, Teevrat, Stuart E. Hamilton, Jacob P. Hochard, Evan Plous Kresch, and John Talbot**, "(Not so) gently down the stream: River pollution and health in Indonesia," *Journal of Environmental Economics and Management*, 2018, *92*, 35–53.

**Hu, Xindi C, David Q Andrews, Andrew B Lindstrom, Thomas A Bruton, Laurel A Schaider, Philippe Grandjean, Rainer Lohmann, Courtney C Carignan, Arlene Blum, Simona A Balan et al.**, "Detection of poly-and perfluoroalkyl substances (PFASs) in US drinking water linked to industrial sites, military fire training areas, and wastewater treatment plants," *Environmental science & technology letters*, 2016, *3* (10), 344–350.

**International Energy Agency**, "Energy and AI: What the data-centre and AI boom could mean for the energy sector," https://www.iea.org/commentaries/energy-and-ai 2024.

_ , "Energy and AI: What the data-centre and AI boom could mean for the energy sector," 2024. Commentary, IEA.

**Jacqz, Irene, Tugba Somuncu, and John Voorheis**, "Fighting Fire with Fire(fighting Foam): The Long Run Effects of PFAS Use at U.S. Military Installations," CES Working Paper 24-72, U.S. Census Bureau, Center for Economic Studies December 2024.

**Jones, Benjamin A.**, "Infant health impacts of freshwater algal blooms: Evidence from an invasive species natural experiment," *Journal of Environmental Economics and Management*, 2019, *96*, 36–59.

**Keiser, David A and Joseph S Shapiro**, "Consequences of the Clean Water Act and the demand for water quality," *The Quarterly Journal of Economics*, 2019, *134* (1), 349–396.

**Knittel, Christopher R, Juan Ramon L Senga, and Shen Wang**, "Flexible data centers and the grid: Lower costs, higher emissions?," Technical Report, National Bureau of Economic Research 2025.

**Lai, Wangyang**, "Pesticide use and health outcomes: Evidence from agricultural water pollution in China," *Journal of Environmental Economics and Management*, 2017, *86*, 93–120. Special issue on environmental economics in developing countries.

**Liang, Yong, Mengshi Lu, Zuo-Jun Max Shen, and Runyu Tang**, "Data Center Network Design for Internet-Related Services and Cloud Computing," *Production and Operations Management*, 2021, *30* (7), 2077–2101.

**Marcus, Michelle and Rosie Mueller**, "Unregulated contaminants in drinking water: Evidence from PFAS and housing prices," *Journal of Environmental Economics and Management*, 2024, *125*, 102987.

**Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey**, "Recalibrating global data center energy-use estimates," *Science*, 2020, *367* (6481), 984–986.

**Rappazzo, Kristen M, Evan Coffman, and Erin P Hines**, "Exposure to perfluorinated alkyl substances and health outcomes in children: a systematic review of the epidemiologic literature," *International journal of environmental research and public health*, 2017, *14* (7), 691.

**Shawhan, Daniel, McKenna Peplinski, Sally Robson, Ethan Russell, Ethan Ziegler, and Karen Palmer**, "Power Delayed: Economic Effects of Electricity Transmission and Generation Development Delays," Technical Report, Resources for the Future 2025.

**Wang, Fengjuan and Chengwei Lv**, "A data center expansion scheme considering net-zero carbon operation: Optimization of geographical location, on-site renewable utilization and green certificate purchase," *Socio-Economic Planning Sciences*, 2024, *92*, 101789.

Figure 1: Spatial distribution of data centers across the U.S.

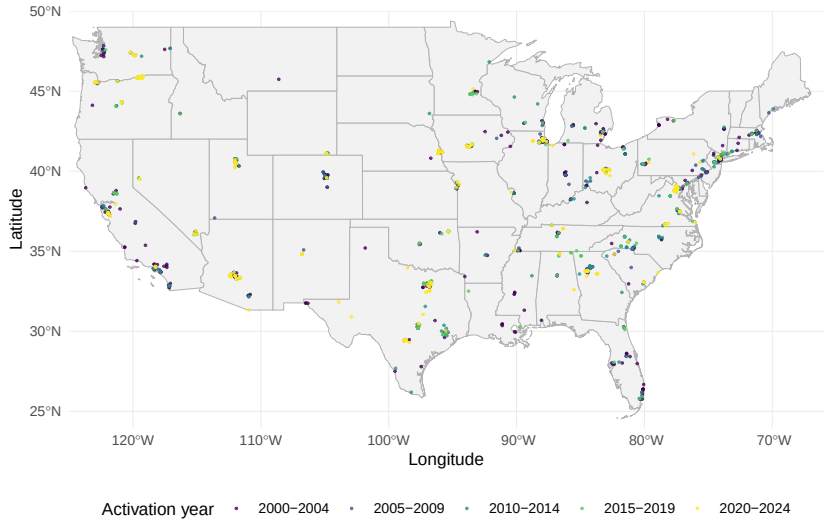
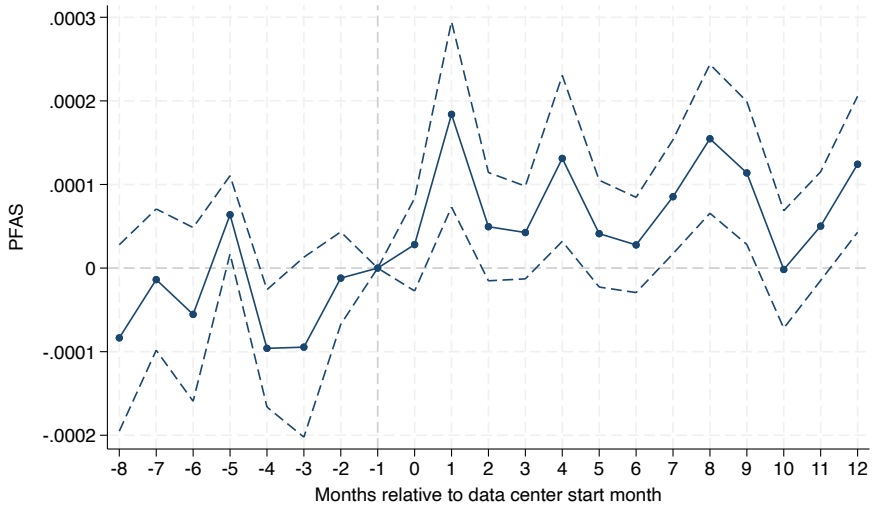



Figure 2: Event study: PFAS concentration around data center openings




Notes: Outcome variable is PFAS concentration with missing values filled using half of the minimum reporting level (MRL). The analysis is conducted at the PWS-zipcode-pollutant-day level for the period January 3, 2023 to June 25, 2025. We include contaminant, PWS, zipcode, year, month, and day-of-week fixed effects, and cluster standard errors at the PWS-zipcode level. Coefficients are normalized relative to the month of data center opening.

Table 1: Effects of data centers on PFAS contamination

| | | | |
|---|---|---|---|
| | Panel A: PFAS concentration ($\mu$g/L) zeros for non-detects | | |
| #Centers (1000) | 0.0222496*** | 0.0238140*** | 0.0182441*** |
| | (0.0077526) | (0.0089790) | (0.0062237) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0734917 | 0.0734918 | 0.0783161 |
| Y-mean | 0.0001726 | 0.0001726 | 0.0001726 |
| Y-sd | 0.0021251 | 0.0021251 | 0.0021251 |
| | Panel B: PFAS concentration ($\mu$g/L) half MRL for non-detects | | |
| #Centers (1000) | 0.0187353*** | 0.0200526*** | 0.0147594*** |
| | (0.0066767) | (0.0077246) | (0.0052856) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.4195757 | 0.4195757 | 0.4224710 |
| Y-mean | 0.0023864 | 0.0023864 | 0.0023864 |
| Y-sd | 0.0025613 | 0.0025613 | 0.0025613 |
| | Panel C: PFAS detection indicator | | |
| #Centers (1000) | 2.0929058*** | 2.2400618*** | 2.1266850*** |
| | (0.6709701) | (0.7721189) | (0.6226139) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.1861949 | 0.1861951 | 0.1909506 |
| Y-mean | 0.0229568 | 0.0229568 | 0.0229568 |
| Y-sd | 0.1371896 | 0.1371896 | 0.1371896 |
| | Panel D: PFAS concentration ($\mu$g/L) zeros for non-detects | | |
| Capacity (TW) | 0.3826360*** | 0.4464955*** | 0.3568448*** |
| | (0.1295399) | (0.1619820) | (0.1354928) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0734914 | 0.0734915 | 0.0783160 |
| Y-mean | 0.0001726 | 0.0001726 | 0.0001726 |
| Y-sd | 0.0021251 | 0.0021251 | 0.0021251 |
| | Panel E: PFAS concentration ($\mu$g/L) half MRL for non-detects | | |
| Capacity (TW) | 0.3264880*** | 0.3809768*** | 0.2921085** |
| | (0.1128156) | (0.1414874) | (0.1168163) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.4195756 | 0.4195756 | 0.4224709 |
| Y-mean | 0.0023864 | 0.0023864 | 0.0023864 |
| Y-sd | 0.0025613 | 0.0025613 | 0.0025613 |
| | Panel F: PFAS detection indicator | | |
| Capacity (TW) | 33.7293676*** | 39.3585843*** | 39.7772663*** |
| | (11.8497263) | (14.1428189) | (13.3987255) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.1861942 | 0.1861944 | 0.1909501 |
| Y-mean | 0.0229568 | 0.0229568 | 0.0229568 |

| Y-sd | 0.1371896 | 0.1371896 | 0.1371896 |
|---|---|---|---|
| PWS FEs | Y | | |
| Zipcode FEs | Y | | |
| PWS-zipcode FEs | | Y | Y |
| Year, month, dow FEs | Y | Y | |
| Date FEs | | | Y |

Notes: This table reports the effects of data center activity on PFAS contamination at the PWS-zipcode-pollutant-day level, January 3, 2023 to June 25, 2025. Panels A-C use the number of active data centers as the exposure measure, while Panels D-F replace center counts with total installed capacity in megawatts (MW). Panels A and D use PFAS concentration with missing values filled with zero, Panels B and E replace missing values with half of the minimum reporting level (MRL), and Panels C and F use a detection indicator. All specifications include combinations of PWS, zipcode, PWS-zipcode, and date fixed effects as shown. Standard errors are clustered at the PWS-zipcode level.

Table 2: Effects of upstream and downstream data centers

| | PFAS concentration ($\mu$g/L) (missing values filled with zero) | PFAS concentration ($\mu$g/L) (missing values filled with half of MRL) | PFAS detection dummy |
|---|---|---|---|
| **Panel A: Upstream data center count** | | | |
| #UpstreamCenters (1000) | 0.0873244** | 0.0793916** | 5.1327811** |
| | (0.0391722) | (0.0350722) | (2.5689004) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0783165 | 0.4224713 | 0.1909497 |
| Y-mean | 0.0001726 | 0.0023864 | 0.0229568 |
| Y-sd | 0.0021251 | 0.0025613 | 0.1371896 |
| **Panel B: Upstream data center capacity** | | | |
| Capacity (TW) | 0.3568448*** | 0.2921085** | 39.7772663*** |
| | (0.1354928) | (0.1168163) | (13.3987255) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0783160 | 0.4224709 | 0.1909501 |
| Y-mean | 0.0001726 | 0.0023864 | 0.0229568 |
| Y-sd | 0.0021251 | 0.0025613 | 0.1371896 |
| **Panel C: Upstream data center count, controlling for same-zipcode data center count** | | | |
| #UpstreamCenters (1000) | 0.0828330** | 0.0758091** | 4.5797701* |
| | (0.0383321) | (0.0344162) | (2.4591728) |
| #Centers (1000) | 0.0164806*** | 0.0131454*** | 2.0291821*** |
| | (0.0053979) | (0.0045265) | (0.5847503) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0783169 | 0.4224714 | 0.1909511 |
| Y-mean | 0.0001726 | 0.0023864 | 0.0229568 |
| Y-sd | 0.0021251 | 0.0025613 | 0.1371896 |
| **Panel D: Upstream data center capacity, controlling for same-zipcode data center capacity** | | | |
| UpstreamCapacity (TW) | 1.4336514** | 1.2991706** | 85.6834854* |
| | (0.7023125) | (0.6298703) | (44.9876691) |
| Capacity (TW) | 0.3129018** | 0.2522875** | 37.1509723*** |
| | (0.1214735) | (0.1040275) | (12.8172620) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0783167 | 0.4224713 | 0.1909506 |
| Y-mean | 0.0001726 | 0.0023864 | 0.0229568 |
| Y-sd | 0.0021251 | 0.0025613 | 0.1371896 |

| | Panel E: Downstream data center count | | |
|---|---|---|---|
| | PFAS concentration ($\mu$g/L) | | PFAS detection dummy |
| | (missing values filled with zero) | (missing values filled with half of MRL) | |
| #DownstreamCenters (1000) | -0.0141081 | -0.0105501 | -1.6602574 |
| | (0.0195982) | (0.0161030) | (2.5760416) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0783157 | 0.4224708 | 0.1909491 |
| Y-mean | 0.0001726 | 0.0023864 | 0.0229568 |
| Y-sd | 0.0021251 | 0.0025613 | 0.1371896 |
| | Panel F: Downstream data center capacity | | |
| | PFAS concentration ($\mu$g/L) | | PFAS detection dummy |
| | (missing values filled with zero) | (missing values filled with half of MRL) | |
| DownstreamCapacity (TW) | -0.0604887 | -0.0281386 | -1.057e+01 |
| | (0.3310522) | (0.2717334) | (43.5561437) |
| Observations | 4009708 | 4009708 | 4009708 |
| R-square | 0.0783157 | 0.4224708 | 0.1909490 |
| Y-mean | 0.0001726 | 0.0023864 | 0.0229568 |
| Y-sd | 0.0021251 | 0.0025613 | 0.1371896 |
| PWS-zipcode FEs | Y | Y | Y |
| Date FEs | Y | Y | Y |

Notes: This table reports the effects of upstream or downstream data center activity on PFAS contamination at the PWS-zipcode-pollutant-day level from January 3, 2023 to June 25, 2025. Panel A measures exposure using the number of active data centers located in upstream zipcodes, while Panel B replaces counts with total installed capacity (MW). Panels C and D additionally control for data centers located in the same zipcode. Panels E and F instead measure exposure using downstream data center counts and capacity. For outcome variables, Column (1) uses PFAS concentration with missing values imputed as zero, Column (2) imputes missing values as one-half of the minimum reporting level (MRL), and Column (3) uses a detection indicator. All specifications include PWS-zipcode and date fixed effects as indicated. Standard errors are clustered at the PWS-zipcode level.

# Online Appendix

## A Appendix figures

Figure S1: Time series of data center construction, 2000-2025

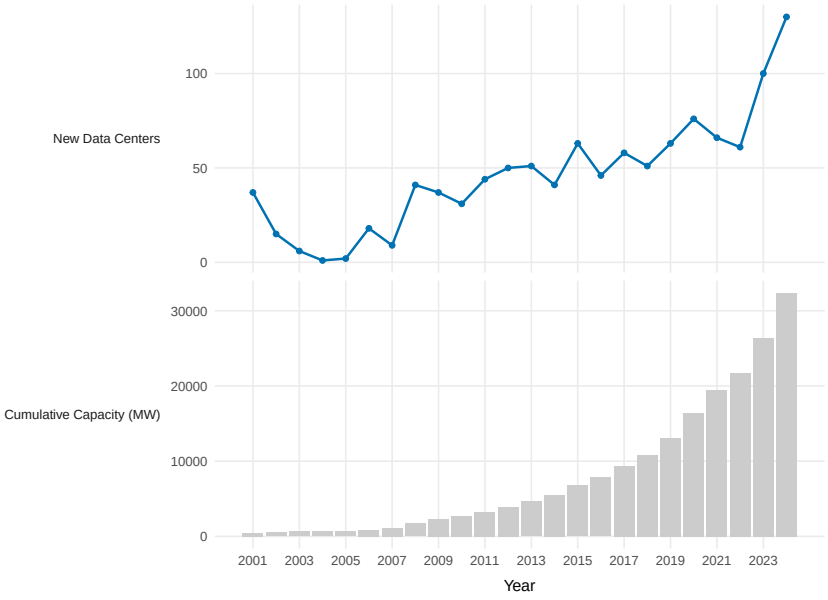
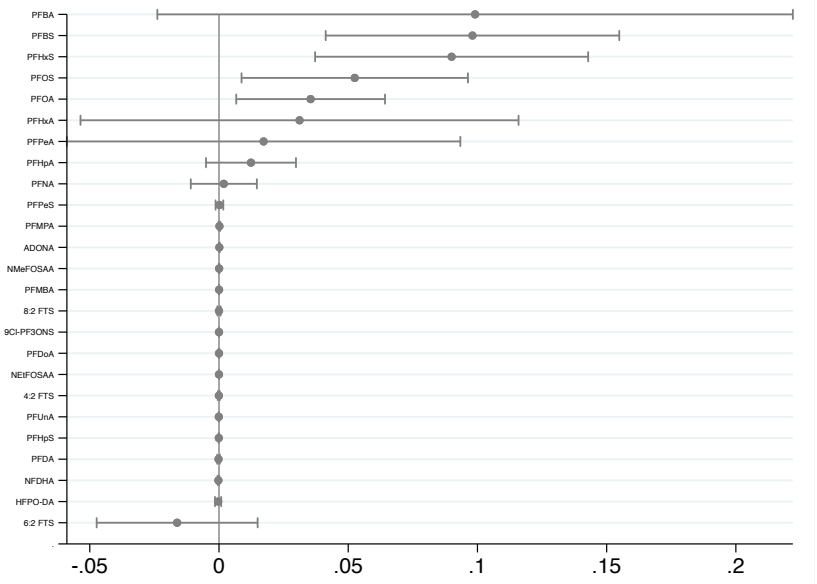
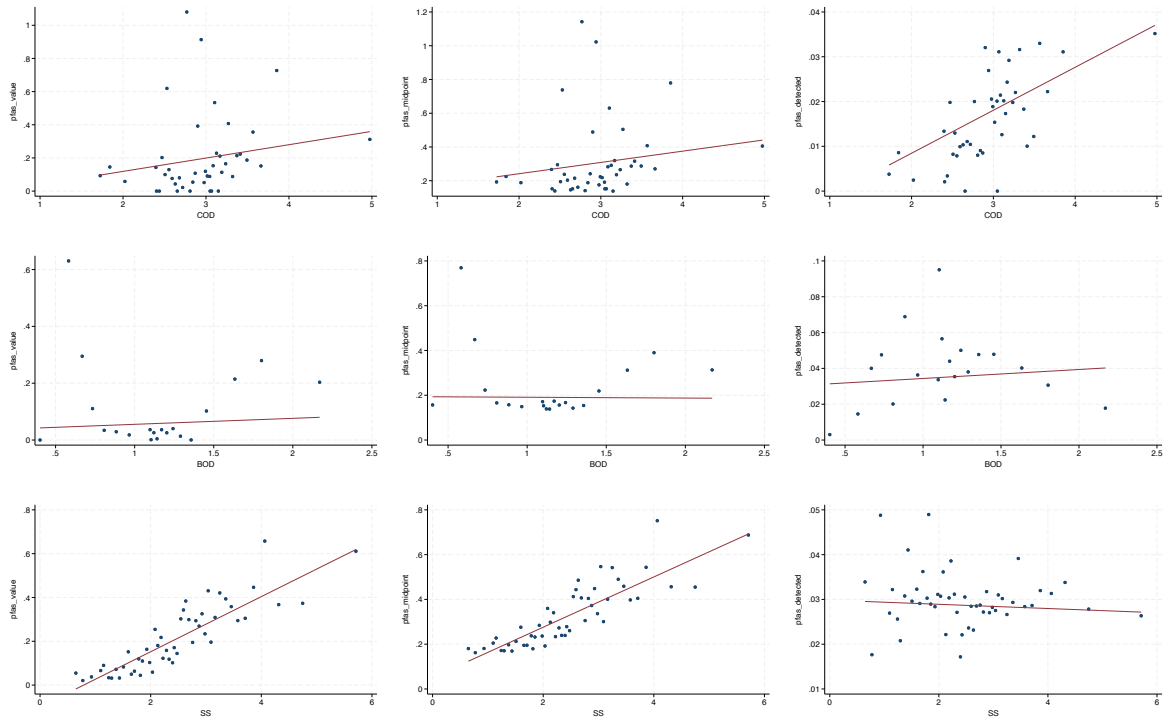



Figure S2: Heterogeneity across PFAS chemicals



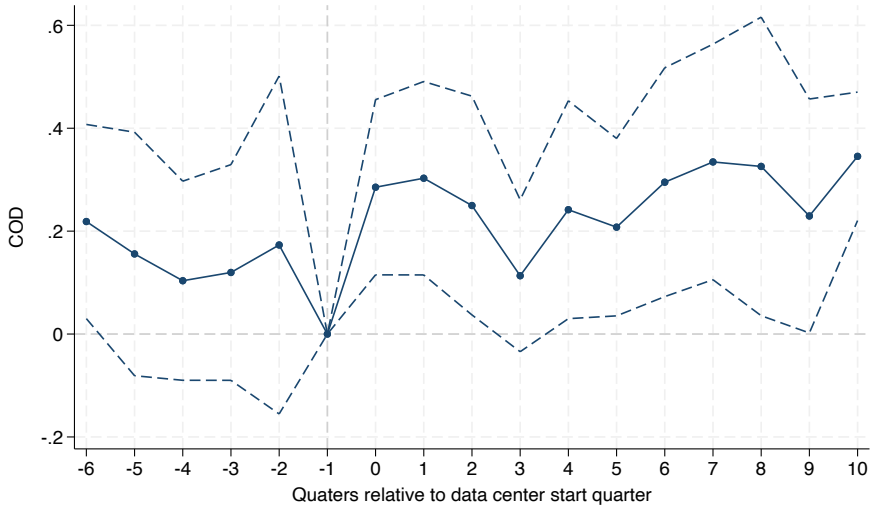

Notes: The analysis is conducted at the PWS-zipcode-pollutant-day level, January 3, 2023 to June 25, 2025. Regressions are estimated separately for each PFAS compound. All specifications include PWS-zipcode and date fixed effects. Standard errors are clustered at the PWS-zipcode level.

Figure S3: Correlation between PFAS and other water pollutants




Notes: The sample is at the zipcode-month level, January 2023 to June 2025. Each panel plots raw data without fixed effects or controls. The x-axis shows the logarithm of COD, BOD, or suspended solids concentration. The y-axis shows PFAS concentration (zeros), PFAS concentration (half MRL), or the PFAS detection indicator.

Figure S4: Event study: COD response to data center openings




Notes: Outcome variable is the logarithm of COD concentration. The analysis is conducted at the zipcode-month level, January 2001 to June 2025. We include zipcode and year-month fixed effects and cluster standard errors at the zipcode level. Coefficients are normalized relative to the month of data center opening.

# B Appendix tables

Table S1: Heterogeneity by on-site power generation

| | Panel E: PFAS concentration ($\mu$g/L) half MRL for non-detects | | | |
|---|---|---|---|---|
| #Centers w/ generator (1000) | 0.0280238*** | | 0.0294005*** | |
| | (0.0093615) | | (0.0113040) | |
| #Centers w/o generator (1000) | | 0.0140295 | -0.0031929 | |
| | | (0.0085506) | (0.0130960) | |
| #Centers (1000) | | | | 0.0155213** |
| | | | | (0.0067140) |
| #Centers (1000) × #Generators (1000) | | | | -0.0046972 |
| | | | | (0.0188741) |
| Observations | 4009708 | 4009708 | 4009708 | 4009708 |
| R-square | 0.4224711 | 0.4224708 | 0.4224711 | 0.4224710 |
| Y-mean | 0.0023864 | 0.0023864 | 0.0023864 | 0.0023864 |
| Y-sd | 0.0025613 | 0.0025613 | 0.0025613 | 0.0025613 |
| PWS-zipcode FEs | Y | Y | Y | Y |
| Date FEs | Y | Y | Y | Y |

Notes: The analysis is conducted at the PWS-zipcode-pollutant-day level, January 3, 2023 to June 25, 2025. We interact data center exposure with an indicator for on-site power generation. All specifications include PWS-zipcode and date fixed effects. Standard errors are clustered at the PWS-zipcode level.

Table S2: Effects of data centers on water pollution

| | | | | |
|---|---|---|---|---|
| | Panel A: COD (mg/L) | | | |
| #Centers | 1.1482111* | 1.0438327* | | |
| | (0.6511646) | (0.6061970) | | |
| Capacity (GW) | | | 40.7228387*** | 31.9449370*** |
| | | | (12.6243693) | (7.9256238) |
| Observations | 85013 | 85013 | 85013 | 85013 |
| R-square | 0.3620903 | 0.3647449 | 0.3620900 | 0.3647445 |
| Y-mean | 2.48e+01 | 2.48e+01 | 2.48e+01 | 2.48e+01 |
| Y-sd | 1.00e+02 | 1.00e+02 | 1.00e+02 | 1.00e+02 |
| | ln(COD+1) | | | |
| #Centers | 0.0721771** | 0.0720374** | | |
| | (0.0328343) | (0.0326190) | | |
| Capacity (GW) | | | 3.7659732*** | 3.7587707*** |
| | | | (0.1712498) | (0.1743399) |
| Observations | 84991 | 84991 | 84991 | 84991 |
| R-square | 0.6293873 | 0.6384350 | 0.6294228 | 0.6384702 |
| | Panel B: BOD (mg/L) | | | |
| #Centers | 0.1413753 | 0.1432914 | | |
| | (0.1321741) | (0.1339624) | | |
| Capacity (GW) | | | 2.7874824 | 2.7631291 |
| | | | (2.5589579) | (2.5730902) |
| Observations | 213116 | 213116 | 213116 | 213116 |
| R-square | 0.1569892 | 0.1578442 | 0.1569890 | 0.1578440 |
| Y-mean | 2.7246147 | 2.7246147 | 2.7246147 | 2.7246147 |
| Y-sd | 4.01e+01 | 4.01e+01 | 4.01e+01 | 4.01e+01 |
| | ln(BOD+1) | | | |
| #Centers | 0.0046210 | 0.0047551 | | |
| | (0.0034881) | (0.0034336) | | |
| Capacity (GW) | | | 0.1058945 | 0.1095100* |
| | | | (0.0665502) | (0.0652414) |
| Observations | 212877 | 212877 | 212877 | 212877 |
| R-square | 0.4829305 | 0.4889982 | 0.4829305 | 0.4889982 |
| | Panel C: Suspended solids (mg/L) | | | |
| #Centers | -0.4387552 | -0.4976467 | | |
| | (0.5272101) | (0.5324666) | | |
| Capacity (GW) | | | -1.177e+01 | -1.340e+01 |
| | | | (17.3069563) | (17.4547345) |
| Observations | 861737 | 861737 | 861737 | 861737 |
| R-square | 0.1278448 | 0.1283055 | 0.1278448 | 0.1283054 |
| Y-mean | 4.69e+01 | 4.69e+01 | 4.69e+01 | 4.69e+01 |
| Y-sd | 6.98e+02 | 6.98e+02 | 6.98e+02 | 6.98e+02 |
| | ln(SS+1) | | | |
| #Centers | -0.0061501 | -0.0060646 | | |
| | (0.0074739) | (0.0075284) | | |
| Capacity (GW) | | | -0.1637611 | -0.1623021 |

| | | | | |
|---|---|---|---|---|
| | | | (0.1916290) | (0.1924635) |
| Observations | 861458 | 861458 | 861458 | 861458 |
| R-square | 0.4985565 | 0.5031375 | 0.4985559 | 0.5031369 |
| Zipcode FEs | Y | Y | Y | Y |
| Year, month FEs | Y | | Y | |
| Year-month FEs | | Y | | Y |

Notes: This table reports the effects of data center activity on water pollution at the zipcode-month level, January 2001 to June 2025. Panels A-C correspond to chemical oxygen demand (COD), biochemical oxygen demand (BOD), and suspended solids (SS), respectively. Each panel reports results in both level and logarithmic specifications. All regressions include zipcode fixed effects and either separate year and month fixed effects or combined year-month fixed effects as indicated. Standard errors are clustered at the zipcode level.

Table S3: Effects of data center openings on labor outcomes

| | | | | |
|---|---|---|---|---|
| | Panel A: ln(Number of workers+1) | | | |
| #Centers | 0.0310*** | 0.0314*** | | |
| | (0.0108) | (0.0108) | | |
| Capacity (GW) | | | 0.6565*** | 0.6683*** |
| | | | (0.2256) | (0.2259) |
| Observations | 24388923 | 24388923 | 24388923 | 24388923 |
| R-square | 0.7652305 | 0.7753805 | 0.7652320 | 0.7753822 |
| Y-mean | 4.4355801 | 4.4355801 | 4.4355801 | 4.4355801 |
| Y-sd | 1.2288830 | 1.2288830 | 1.2288830 | 1.2288830 |
| | Panel B: ln(Working hours per worker+1) | | | |
| #Centers | 0.0019 | 0.0019 | | |
| | (0.0014) | (0.0014) | | |
| Capacity (GW) | | | 0.0364 | 0.0361 |
| | | | (0.0287) | (0.0287) |
| Observations | 24388923 | 24388923 | 24388923 | 24388923 |
| R-square | 0.3117427 | 0.3158202 | 0.3117425 | 0.3158200 |
| Y-mean | 2.0397393 | 2.0397393 | 2.0397393 | 2.0397393 |
| Y-sd | 0.2490082 | 0.2490082 | 0.2490082 | 0.2490082 |
| | Panel C: ln(Total earnings per worker+1) | | | |
| #Centers | 0.0142 | 0.0142 | | |
| | (0.0190) | (0.0190) | | |
| Capacity (GW) | | | 0.7955* | 0.7982* |
| | | | (0.4662) | (0.4673) |
| Observations | 24388923 | 24388923 | 24388923 | 24388923 |
| R-square | 0.5839864 | 0.5856114 | 0.5840001 | 0.5856252 |
| Y-mean | 3.3027278 | 3.3027278 | 3.3027278 | 3.3027278 |
| Y-sd | 1.6086912 | 1.6086912 | 1.6086912 | 1.6086912 |
| | Panel D: ln(Hourly wage+1) | | | |
| #Centers | 0.0092 | 0.0093 | | |
| | (0.0126) | (0.0126) | | |
| Capacity (GW) | | | 0.4800 | 0.4820 |
| | | | (0.3047) | (0.3053) |
| Observations | 24370870 | 24370870 | 24370870 | 24370870 |
| R-square | 0.5871610 | 0.5888096 | 0.5871746 | 0.5888232 |
| Y-mean | 1.8265235 | 1.8265235 | 1.8265235 | 1.8265235 |
| Y-sd | 0.9652110 | 0.9652110 | 0.9652110 | 0.9652110 |
| Zipcode FEs | Y | Y | Y | Y |
| Year, month, dow FEs | Y | | Y | |
| Date FEs | | Y | | Y |

Notes: The analysis is conducted at the industry-zipcode-day level from June 28, 2020 to October 28, 2025. The key explanatory variables are the number of active data centers and total data center capacity (GW). Standard errors are clustered at the zipcode level.

Table S4: Heterogeneity by worker level

| | Panel A: ln(Number of workers+1) | | |
|---|---|---|---|
| | Employee | Manager | General manager |
| #Centers | 0.0289*** | 0.0197 | 0.0218 |
| | (0.0100) | (0.0253) | (0.0207) |
| Observations | 24388923 | 24388923 | 24388923 |
| R-square | 0.7482832 | 0.6002786 | 0.4551290 |
| Y-mean | 4.3362736 | 1.3763768 | 0.3589909 |
| Y-sd | 1.2628607 | 1.4847851 | 0.8762462 |
| | Panel B: ln(Working hours per worker+1) | | |
| | Employee | Manager | General manager |
| #Centers | 0.0020 | -0.0017 | -0.0113 |
| | (0.0014) | (0.0027) | (0.0081) |
| Observations | 24172605 | 12032154 | 3659878 |
| R-square | 0.3215091 | 0.1965331 | 0.2125962 |
| Y-mean | 2.0356270 | 2.0885962 | 2.1061034 |
| Y-sd | 0.2493003 | 0.3567985 | 0.4036417 |
| | Panel C: ln(Total earnings per worker+1) | | |
| | Employee | Manager | General manager |
| #Centers | 0.0165 | -0.0068 | 0.0153 |
| | (0.0188) | (0.0306) | (0.0714) |
| Observations | 24172605 | 12032154 | 3659878 |
| R-square | 0.5822961 | 0.4843291 | 0.5362649 |
| Y-mean | 3.2881532 | 3.1043345 | 2.6651453 |
| Y-sd | 1.6107673 | 2.0891545 | 2.3328264 |
| | Panel D: ln(Hourly wage+1) | | |
| | Employee | Manager | General manager |
| #Centers | 0.0106 | -0.0029 | 0.0082 |
| | (0.0124) | (0.0193) | (0.0446) |
| Observations | 24155857 | 12007428 | 3651660 |
| R-square | 0.5864271 | 0.4958758 | 0.5548858 |
| Y-mean | 1.8203296 | 1.7934405 | 1.5896023 |
| Y-sd | 0.9675673 | 1.2423884 | 1.4087329 |
| Zipcode FEs | Y | Y | Y |
| Date FEs | Y | Y | Y |

Notes: The analysis is conducted at the zipcode-day level from June 28, 2020 to October 28, 2025. Regressions are estimated separately for each worker level. Standard errors are clustered at the zipcode level.

Table S5: Heterogeneity by industry

| | Panel A: ln(Number of workers+1) | | | | |
|---|---|---|---|---|---|
| | Beauty & Personal care | Charities & Education | Food & Drink | Healthcare & Fitness | Home & Repair |
| #Centers | 0.0001 | -0.0047 | 0.0306*** | 0.0128 | 0.0238 |
| | (0.0036) | (0.0109) | (0.0107) | (0.0135) | (0.0176) |
| Observations | 408930 | 3559956 | 17947277 | 7110325 | 2845636 |
| R-square | 0.6682644 | 0.5891536 | 0.7429578 | 0.6071769 | 0.6117726 |
| Y-mean | 2.8535987 | 3.5583847 | 4.2550048 | 3.5248718 | 3.1726125 |
| Y-sd | 0.7269459 | 1.0199315 | 1.0816660 | 0.9547609 | 0.8759931 |
| | Leisure & Entertainment | Professional service | Retail | Transportation | |
| #Centers | 0.0226*** | 0.0247 | -0.0124 | 0.0667*** | |
| | (0.0081) | (0.0177) | (0.0205) | (0.0185) | |
| Observations | 2446140 | 5099182 | 11566435 | 1456312 | |
| R-square | 0.6304632 | 0.6162690 | 0.6741689 | 0.7078247 | |
| Y-mean | 3.5227995 | 3.2975450 | 3.5984274 | 3.5012764 | |
| Y-sd | 0.9788581 | 0.9592922 | 0.9712687 | 1.0398915 | |
| Zipcode FEs | Y | Y | Y | Y | Y |
| Year, month, dow FEs | Y | | Y | | |
| Date FEs | | Y | | Y | |

Notes: The analysis is conducted at the zipcode-day level from June 28, 2020 to October 28, 2025. The dependent variable is ln(Number of workers+1) and regressions are estimated separately by industry. Standard errors are clustered at the zipcode level.