

# The Unpaid Toll: Quantifying and Addressing the Public Health Impact of Data Centers

Yuelin Han  
*UC Riverside*

Zhifeng Wu  
*UC Riverside*

Pengfei Li  
*RIT*

Adam Wierman  
*Caltech*

Shaolei Ren<sup>1</sup>  
*UC Riverside*

## Abstract

The surging demand for AI has led to a rapid expansion of energy-intensive data centers, impacting the environment through escalating carbon emissions and water consumption. While significant attention has been paid to data centers' growing environmental footprint, the public health burden, a hidden toll of data centers, has been largely overlooked. Specifically, data centers' lifecycle, from chip manufacturing to operation, can significantly degrade air quality through emissions of criteria air pollutants such as fine particulate matter, substantially impacting public health. This paper introduces a principled methodology to model life-cycle pollutant emissions for data centers and computing tasks, quantifying the public health impacts. Our findings reveal that training a large AI model comparable to the Llama-3.1 scale can produce air pollutants equivalent to more than 10,000 round trips by car between Los Angeles and New York City. The growing demand for AI is projected to push the total annual public health burden of U.S. data centers up to more than \$20 billion in 2028, rivaling that of on-road emissions of California. Further, the public health costs are more felt in disadvantaged communities, where the per-household health burden could be 200x more than that in less-impacted communities. Finally, we propose a health-informed computing framework that explicitly incorporates public health risk as a key metric for scheduling data center workloads across space and time, which can effectively mitigate adverse health impacts while advancing environmental sustainability. More broadly, we also recommend adopting a standard reporting protocol for the public health impacts of data centers and paying attention to all impacted communities.

## 1 Introduction

The rise of artificial intelligence (AI) has numerous potentials to play a transformative role in addressing grand societal challenges, including air quality and public health [1, 2]. For example, by integrating multimodal data from various sources, AI can provide effective tools and actionable insights for pandemic preparedness, disease prevention, healthcare optimization, and air quality management [1, 3]. However, the surging demand for AI — particularly generative AI, as exemplified by the recent popularity of large language models (LLMs) — has driven a rapid increase in computational needs, fueling the unprecedented expansion of energy-intensive data centers. According to the recent Lawrence Berkeley National Lab report [4], AI training and inference are projected to become the dominant workloads and push the U.S. data center electricity consumption to account for 6.7–12.0% of the national total in 2028, up from 4.4% in 2023.

The growing electricity demand of data centers has not only created significant stress on power grid stability [5, 6], but also increasingly impacts the environment through escalating carbon emissions [7, 8] and water consumption [9]. These environmental impacts are driven primarily by the “expansion of AI products and services,” as recently acknowledged by technology companies in their sustainability reports [10]. To mitigate the challenges posed to both power grids and the environment, a range of strategies have been explored, including grid-integrated data centers [6, 11], energy-efficient hardware and software [12–14], and the adoption of carbon-aware and water-efficient computing practices [9, 15–17], among others.

**The hidden toll of data centers.** While the environmental footprint of data centers has garnered attention, the public health burden, a hidden toll of data centers, has been largely overlooked. Across its entire lifecycle — from chip manufacturing to operation — a data center contributes substantially to air quality degradation and public health costs through the emission of various criteria air pollutants. These include fine particulate matter ( $PM_{2.5}$ , particles measuring 2.5 micrometers or smaller in diameter that can penetrate

---

<sup>1</sup> Yuelin Han and Zhifeng Wu contributed equally and are listed alphabetically.

Corresponding authors: Adam Wierman ([adamw@caltech.edu](mailto:adamw@caltech.edu)) and Shaolei Ren ([shaolei@ucr.edu](mailto:shaolei@ucr.edu))

deep into lungs and cause serious health effects), sulfur dioxide ( $\text{SO}_2$ ), and nitrogen dioxide ( $\text{NO}_2$ ). Concretely, the server manufacturing process [18], electricity generation from fossil fuels to power data centers, and the maintenance and usage of diesel backup generators to ensure continuous data center operation all produce significant amounts of criteria air pollutants. Moreover, the distinct spatial-temporal heterogeneities of emission sources suggest that focusing solely on reducing data centers' carbon footprints may not minimize its emissions of criteria air pollutants or the resulting public health impacts (Section 6.2).

Exposure to criteria air pollutants is directly and causally linked to various adverse health outcomes,<sup>2</sup> including premature mortality, lung cancer, asthma, heart attacks, cardiovascular diseases, and even cognitive decline, especially for the elderly and vulnerable individuals with pre-existing conditions [20–22]. Moreover, even short-term (hours to days)  $\text{PM}_{2.5}$  exposure is harmful and deadly, accounting for approximately 1 million premature deaths per year from 2000 to 2019 and representing 2% of total global deaths [23].

Criteria air pollutants are not confined to the immediate vicinity of their emission sources; they can travel hundreds of miles through a dispersion process (i.e., cross-state air pollution) [24, 25], impacting public health across vast regions. Further,  $\text{PM}_{2.5}$  is considered “non-threshold,” i.e., there is no absolutely safe exposure level [26]. Thus, compliance with the national/regional air quality standards does not necessarily ensure the air is healthy.

Globally, 4.2 million deaths were attributed to ambient (i.e., outdoor) air pollution in 2019 [27]. Air pollution has become the second highest risk factor for noncommunicable diseases [28]. Notably, according to the latest Global Burden of Disease report [29], along with high blood pressure and high blood sugar, ambient particulate matter is placed among the leading risk factors for disease burden globally in every socio-demographic group.

Importantly, along with transportation and industrial activities, electricity generation is a key contributor to ambient air pollution with substantial public health impacts [28, 30, 31]. For example, a recent study [32] shows that, between 1999 and 2020, a total of 460,000 excess deaths were attributed to  $\text{PM}_{2.5}$  generated by coal-fired power plants alone in the U.S. As highlighted by the U.S. EPA [30], despite years of progress, power plants “remain a leading source of air, water, and land pollution that affects communities nationwide.” In Europe, the public health cost of air pollution from power plants is valued at approximately 1% of the gross domestic product (GDP), according to the European Environment Agency’s study in 2024 [33].

The public health outcomes of data centers due to their emission of criteria air pollutants lead to various losses, such as hospitalizations, medication usage, emergency room visits, school loss days, and lost work-days. Nonetheless, despite recent policy efforts [34, 35], the tangible and growing public health impacts of data centers have remained under the radar, almost entirely omitted from today’s risk assessments and sustainability reports [10, 36, 37].

**Quantifying and addressing the public health impacts of data centers.** In this paper, we uncover and quantify the hidden public health impacts of data centers. We introduce a principled methodology to model the emission of criteria air pollutants associated with a computing task and data center across three distinct scopes: emissions from the maintenance and operation of backup generators (Scope 1), emissions from fossil fuel combustion for electricity generation (Scope 2), and emissions resulting from the manufacturing of server hardware (Scope 3). Then, we analyze the dispersion of criteria air pollutants and the resulting public health impacts.

As the U.S. hosts nearly half of the world’s data centers [38] and the EPA data excludes other regions [39], our empirical study focuses on the 48 contiguous U.S. states plus Washington D.C.<sup>3</sup> Our main results (Section 5) focus on the scope-1 and scope-2 health impacts of U.S. data centers and, specifically, LLM training. Using the reduced-complexity modeling tool COBRA (CO-Benefits Risk Assessment) provided by the EPA [39], our analysis demonstrates that driven by the growing demand for AI, the U.S. data centers could contribute to, among others, approximately 600,000 asthma symptom cases and 1,300 premature deaths in 2028, exceeding 1/3 of asthma deaths in the U.S. each year [40]. The overall public health costs could reach more than \$20 billion, rival or even top those of on-road emissions of the largest U.S. states such as California

<sup>2</sup>While we focus on public health, we note that the impacts of criteria air pollutants extend beyond humans and include harms to environmentally sensitive areas, such as some national parks and wilderness areas which, classified as “Class 1 areas” under the Clean Air Act, require special air protection [19].

<sup>3</sup>If located in countries with higher population densities, more pollutant-intensive electricity mixes, or less stringent air quality standards, the same data centers would likely lead to more premature deaths and other adverse health impacts than in the U.S. We recommend further research on the public health impact of non-U.S. data centers.

with ~35 million registered vehicles [41]. Moreover, depending on the location, training an AI model of the Llama-3.1 scale can produce an amount of air pollutants equivalent to driving a car for more than 10,000 round trips between Los Angeles and New York City (LA-NYC), resulting in a health cost that even exceeds 120% of the training electricity cost.

Importantly, although the public health impact of data centers may be modest at the national level, it is geographically concentrated, with certain regions and communities bearing a disproportionate share. In particular, some low-income counties experience significantly greater health costs, with per-household burdens exceeding those in other counties by more than 200-fold.

Furthermore, to highlight scope-1 health impacts, we examine data center backup generators in Virginia, which hosts one of the largest concentrations of data centers in the world [42]. Our analysis shows that, assuming the actual emissions are only 10% of the permitted level based on the historical reports and future projections [42–44], the data center backup generators registered in Virginia (mostly in Loudoun, Prince William, and Fairfax) could already cause 14,000 asthma symptom cases among other health outcomes and a total public health burden of \$220-300 million per year, impacting residents in multiple surrounding states and as far as Florida (Section 3.1). If these data centers emit air pollutants at the maximum permitted level, the total public health cost will become 10-fold and reach \$2.2-3.0 billion per year.

To address the growing and uneven distribution of health burdens, we propose health-informed computing that explicitly incorporates the public health risk as a key metric when scheduling data center workloads. Specifically, by exploiting the spatial-temporal variations of public health impacts and using spatial load shifting as a case study, we demonstrate that the health-informed approach can significantly reduce the health cost compared to the baseline, while continuing to offer meaningful electricity cost savings and reductions in carbon emissions.

Finally, we provide broader recommendations to address the increasing public health impact of data centers (Section 7). We recommend technology companies adopt a standard reporting protocol for criteria air pollutants and public health impacts in their AI model cards and sustainability reports and pay attention to all impacted communities.

To summarize, while AI and data centers offer many societal benefits, our study sheds light on, quantifies, and addresses the often overlooked negative externalities of their resource demand, particularly the public health impact. We also urge further research to comprehensively address the public health implications when developing data centers in the future, ensuring that the growth of data centers does not exacerbate the health burden.

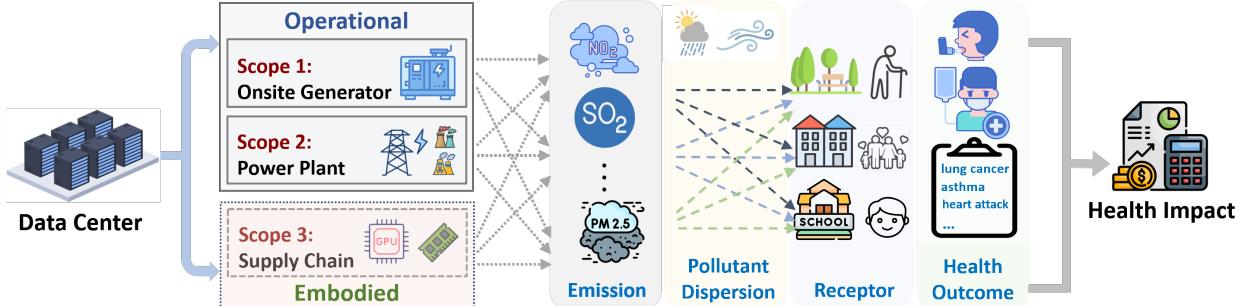
**Disclaimer.** *The results presented in this paper are not intended to encourage or discourage the construction of data centers, nor should they be used to support or oppose any specific project, which requires more detailed and context-specific evaluation. We do not take a position on decisions related to any specific data centers or the use of AI, but instead provide a quantitative assessment of the potential public health impacts of the data center industry.*

## 2 Background on Air Pollutants

This section provides background on criteria air pollutants and U.S. air quality policies. Other countries have similar policies in place to safeguard public health, although their levels of enforcement strictness often differ [33].

Criteria air pollutants, including PM<sub>2.5</sub>, SO<sub>2</sub> and NO<sub>2</sub>, are a group of airborne contaminants that are emitted from various sources such as industrial activities and vehicle emissions. The direct emission of PM<sub>2.5</sub> is called primary PM<sub>2.5</sub>, while precursor pollutants such as SO<sub>2</sub>, NO<sub>x</sub>, and VOCs, can form secondary PM<sub>2.5</sub> and/or ozone [45]. These air pollutants can travel a long distance (a.k.a. cross-state air pollution), posing direct and significant risks to public health over large areas, particularly for vulnerable populations including the elderly and individuals with respiratory conditions [24, 25].

Long-term exposure to PM<sub>2.5</sub>, even at a low level, are directly linked to numerous health outcomes, including premature mortality, heart attacks, asthma, stroke, lung cancer, and even cognitive decline [21, 22]. These health effects result in various losses, such as hospitalizations, medication usage, emergency room visits, school loss days, and lost workdays, which can be further quantified in economic costs based on public health research for various health endpoints [46]. In addition, short-term (hours to days) PM<sub>2.5</sub> exposure is also dangerous, contributing to approximately 1 million premature deaths per year globally from 2000 to 2019 [23].



**Figure 1:** The overview of data centers’ contribution to air pollutants and public health impacts. Scope-1 and scope-2 impacts occur during the operation of data centers (“operational”), whereas scope-3 impacts arise from activities across the supply chain (“embodied”).

Under the Clean Air Act, the U.S. EPA is authorized to regulate the emission levels of criteria air pollutants, reducing concentrations to comply with the National Ambient Air Quality Standards (NAAQS) [47]. For example, the NAAQS primary standards set the annual average PM<sub>2.5</sub> concentration at 9  $\mu\text{g}/\text{m}^3$  and the 98-th percentile of 1-hour daily maximum NO<sub>2</sub> concentration at 100 parts per billion by volume, both counted over three years [48]. In addition, state and local governments may set additional regulations on criteria air pollutants to strengthen or reinforce national standards [49].

While the U.S. has generally better air quality than many other countries, 4 in 10 people in the U.S. still live with unhealthy levels of air pollution, according to the “State of the Air 2024” report published by the American Lung Association [50]. In 2019 (the latest year of data provided by the World Health Organization, or WHO, as of November 2024), an estimate of 93,886 deaths in the U.S. were attributed to ambient air pollution [51]. In fact, the EPA’s recently tightened standard for PM<sub>2.5</sub> sets an annual average limit of 9  $\mu\text{g}/\text{m}^3$ , considerably higher than the WHO’s recommended level of 5  $\mu\text{g}/\text{m}^3$  [48, 52]. Moreover, the EPA projects that 53 U.S. counties, including 23 in the already most populous state of California, would fail to meet the revised national annual PM<sub>2.5</sub> standard in 2032 [53].

Although CO<sub>2</sub> is broadly classified by the EPA as an air pollutant following the U.S. Supreme Court ruling in 2007 [54] and contributes to long-term climate change, it often does not cause the same immediate health impacts as criteria pollutants [55]. In the U.S., CO<sub>2</sub> and other greenhouse gases are subject to different EPA regulations from those for criteria air pollutants. Thus, for the sake of presentation in this paper, we use “air pollutants” to solely refer to criteria air pollutants wherever applicable.

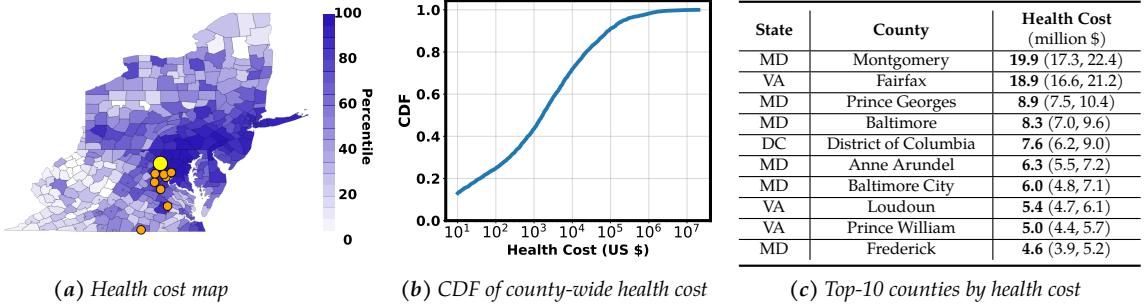
### 3 Data Centers’ Contribution to Air Pollutants

This section presents an overview of data centers’ impact on air quality and contribution to criteria air pollutants throughout its lifecycle across three scopes (Fig. 1). The scoping definition in this paper parallels the well-established greenhouse gas protocol [56]. Specifically, scope-1 and scope-2 air pollutants primarily originate from onsite generators and power plants, collectively referred to as operational emissions, while scope-3 pollutants arise from the supply chain and are referred to as embodied emissions.

#### 3.1 Scope 1: Onsite Generator

While the construction phase of a data center directly increases air pollutant emissions, its amortized health impact over a typical 15–20 year lifespan is negligible. Therefore, we focus on onsite backup generators as the primary source of scope-1 direct air pollutants.

Data centers are mission-critical facilities that are designed to operate with high availability and uptime guarantees. As a result, to maintain operation during emergencies such as grid outages, data centers require highly reliable backup power sources [10, 37]. Diesel generators are known to emit significant amounts of air pollutants and even hazardous emissions during operation [57]. For example, they emit 200–600 times more NO<sub>x</sub> than new or controlled existing natural gas-fired power plants for each unit of electricity produced [58]. Moreover, capacity redundancy is typically followed for diesel generator installations to ensure high availability [59]. Nonetheless, there is limited practical experience with cleaner backup alternatives that



**Figure 2:** The county-level total scope-1 health cost of data center backup generators operated in Virginia (mostly in Loudoun County, Fairfax County, and Prince William County) [62]. The backup generators are assumed to emit air pollutants at 10% of the permitted levels per year. The total annual public health cost is \$220-300 million, including \$190-260 million incurred in Virginia, West Virginia, Maryland, Pennsylvania, New York, New Jersey, Delaware, and Washington D.C. (a) County-level health cost in Virginia, West Virginia, Maryland, Pennsylvania, New York, New Jersey, Delaware, and Washington D.C. Counties with data centers are marked in orange, except for Loudoun County (marked in yellow). (b) CDF of the county-level cost. (c) Top-10 counties by the total health cost.

can provide comparable reliability at scale in real-world settings in the near term, as highlighted by the U.S. Department of Energy in its recent recommendations regarding AI data center infrastructures [6].

Consequently, the vast majority of data centers, even including those newly built by major technology companies, depend on onsite diesel generators for backup power [6,60]. For example, in Northern Virginia (mostly in Loudoun, Prince William, and Fairfax), the number of permits for data center diesel generators has increased by more than 70% since 2023 compared to the total number of permits issued between 2000 and 2022 [60]. Importantly, nearly all the diesel generators are Tier 2, which have significantly higher emission rates than Tier 4 units [60,61]. The total permitted annual emission limits for these diesel generators are approximately 13,000 tons of NO<sub>x</sub>, 1,400 tons of VOCs, 50 tons of SO<sub>2</sub>, and 600 tons of PM<sub>2.5</sub>, all in U.S. short tons.

While diesel generators need to comply with air quality regulations and typically do not operate over extended periods of time, regular maintenance and testing are essential to ensure their operational reliability. A recent report by the state of Virginia [42] found that the actual air pollutant emissions from backup generators at Virginia’s data centers reached approximately 7% of the total permitted amounts in 2023, primarily for maintenance. Likewise, the actual emissions took up 3% to 12% of the permitted levels for some data centers in Quincy, Washington [43].

Moreover, the U.S. EPA recently issued a clarification that would allow data centers to run backup generators for up to 50 hours a year (or roughly 10% of the permits that typically allow 500 hours per year) to participate in demand response—a program designed to reduce grid demand during peak hours, which is increasingly activated as surging data center demand strains grid capacity [44]. This trend may necessitate extended reliance on backup generators [6]. What further adds to the public health impact is that many data center generators in a region may operate simultaneously for demand response during grid capacity shortages, potentially resulting in a short-term spike in PM<sub>2.5</sub> and NO<sub>x</sub> emissions that can be particularly harmful [6,23,48]. For example, from June 23 to 25, 2025, some data centers in Loudoun County, Virginia, were instructed to run their on-site diesel generators for demand response, releasing large amounts of air pollutants and “black smoke” [61].

The high emission rate from onsite generators, combined with extended operation for maintenance and demand response beyond grid outages, could pose serious health risks, especially in regions with a concentration of large data centers. To illustrate this point, we consider the data centers’ onsite generators in Virginia. Assuming that the actual emissions are 10% of the permitted level as a reference case that reflects both the historical reports and future demand response projections [42–44],<sup>4</sup> the backup generators could already cause 14,000 asthma symptom cases and 13–19 deaths each year among other health implica-

<sup>4</sup>If the actual percentage is  $x\%$ , our value will be approximately scaled by  $\frac{x}{10}$  [22].

tions, resulting in a total annual public health burden of \$220-300 million throughout the U.S. This includes \$190-260 million in Virginia, West Virginia, Maryland, Pennsylvania, Delaware, New Jersey, New York, and Washington D.C. We show the county-level health cost and the top-10 counties in Figure 2, while deferring the details of calculations to Appendix A.1. If the diesel generators in Northern Virginia emit air pollutants at the maximum permitted level, the emission of NO<sub>x</sub> could even exceed half of the annual total emissions by all sources in the region [42], resulting in a total public health cost of \$2.2-3.0 billion per year.

### 3.2 Scope 2: Power Plants

Just as data centers are accountable for scope-2 carbon emissions, they also contribute to scope-2 air pollution through their electricity usage.

Along with transportation and manufacturing, the combustion of fossil fuels for electricity production is a leading anthropogenic source of criteria air pollutants, releasing large amounts of PM<sub>2.5</sub>, SO<sub>2</sub>, NO<sub>x</sub>, VOCs, and others [30].<sup>5</sup> More alarmingly, the growing energy demands of AI data centers are already delaying the decommissioning of coal-fired power plants and driving the expansion of fossil-fuel-based plants in the U.S. and globally [6, 65, 66]. For example, in addition to keeping 2,099 MW coal generation capacity until 2039 (more than 80% of the 2024 level), Virginia Electric and Power Company plans to install 5,934 MW gas-fired plants to meet the growing energy demand driven by AI data centers [66].

Based on the emission data projected by the U.S. EPA's COBRA modeling tool [39], we show in Fig. 3 that the electric power sector's total public health cost in the contiguous U.S. is on track to rise, rivaling that of on-road vehicle emissions by all the registered vehicles (including tailpipe exhausts and brakes) in 2028.

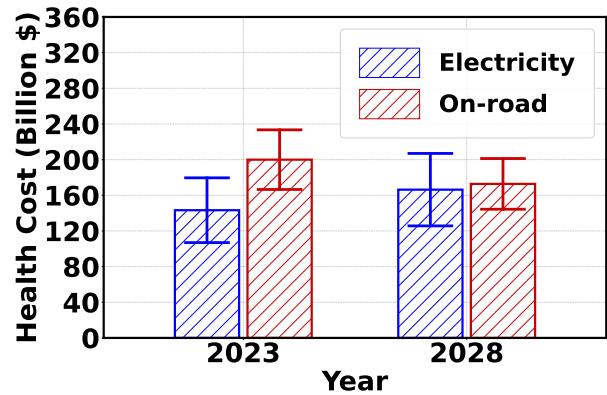
Looking forward, the U.S. Energy Information Administration (EIA) projects that coal consumption by the electricity sector in 2050 will still be approximately 30% of the 2024 level if power plants continue operating under rules existing prior to April 2024 [67]. At the global scale, the reliance on coal and other fossil fuels for electricity production has seen little change over the past four decades, highlighting the enduring challenges to fully adopt clean energy for powering data centers [68]. As a result, data centers' scope-2 air pollution is expected to remain at a high level for a substantially long time into the future.

Although technology companies have started implementing various initiatives—such as purchasing renewable energy credits and nuclear power from small modular reactors [5, 10, 69]—to lower their (market-based) carbon emissions, the vast majority of U.S. data centers remain physically and directly powered by local power grids with a substantial portion of fossil fuel-based energy sources [10]. While the increasingly stringent transmission line constraint is also driving the co-location of data centers with power plants to speed up the construction process, the onsite plants are often gas-fired [70, 71], raising health concerns [31].

We also note that the practice of using various credits to offset scope-2 carbon emissions [10] may not be effective for mitigating the scope-2 public health impact. The reason is that the public health impact of using grid electricity is highly location-dependent, e.g., the impact in a populated region may not be mitigated by renewable energy generated elsewhere.

### 3.3 Scope 3: Supply Chain

The surging demand for AI data centers necessitates large quantities of computational hardware, including graphics processing units (GPUs), thus intensifying the supply chain requirements [72]. However, semiconductor manufacturing generates various criteria air pollutants, wastewater, toxic materials, and hazardous



**Figure 3:** Public health costs of electricity generation and on-road emissions in the contiguous U.S. in 2023 and 2028 [39]. The error bars represent high and low estimates returned by COBRA using two different exposure-response functions.

<sup>5</sup>Wet cooling towers, including those used by data centers [9, 10] and carbon-free nuclear power plants, rely on water evaporation for heat rejection and produce PM<sub>2.5</sub> due to spray drift droplets [63, 64]. Nonetheless, because of limited data available, we exclude the cooling tower PM<sub>2.5</sub> emission from our analysis unless otherwise specified.

air emissions [18]. Moreover, the energy-intensive nature of semiconductor production further contributes to pollutants from power plants. Combined with other pollution sources such as transportation and electronic waste recycling [73], the supply chain activities form a large portion of data centers' scope-3 impact on public health.

Although semiconductor manufacturing facilities are subject to air quality regulations [74], they still pose significant risks, affecting populations across large regions. Importantly, the global demand for AI chips in 2030 is projected to be tens of times of the overall production capacity of this single facility [75], further magnifying the overall scope-3 public health impact of data centers. It is also worth noting that additional pollutants, including hazardous air pollutants like hydrogen fluoride, may further elevate public health costs but are not included in this paper.

However, public data on scope-3 criteria air pollutant emissions from semiconductor manufacturers is limited. Thus, we focus on scope-1 and scope-2 emissions, excluding scope-3 impacts from the main analysis. A detailed assessment of the health impacts associated with a specific U.S. semiconductor manufacturing facility is provided in Appendix A.3.

## 4 Quantifying Task-Specific Public Health Impact

To quantify the public health impact of a specific computing task, we present a principled end-to-end methodology illustrated in Fig. 1. Specifically, the process includes: (1) Quantifying the task's criteria air pollutants at the emission source; (2) Modeling the dispersed air pollutants at different receptors (i.e., destination regions); (3) Calculating the public health impact and assigning a corresponding monetary value at each receptor.

For a computing task under consideration (e.g., AI model training), we consider  $M$  types of criteria air pollutants,  $N$  receptor regions of interest (e.g., all the U.S. counties),  $H$  types of public health impacts (e.g., mortality, asthma symptoms, school loss days, etc.). We use  $p^s = (p_1^s, \dots, p_M^s)$  and  $p_i^r = (p_{i,1}^r, \dots, p_{i,M}^r)$  denote the quantities for  $M$  types of air pollutants attributed to the task at the emission source and at the receptor  $i$ , respectively, for  $i = 1, \dots, N$ . Additionally, we use  $h_i = (h_{i,1}, \dots, h_{i,H})$  and  $c_i = (c_{i,1}, \dots, c_{i,H})$  to denote the incidences and economic costs associated with  $H$  types of health impacts at receptor  $i$ , respectively, for  $i = 1, \dots, N$ . With a slight abuse of notations, we reuse these symbols when modeling AI's public health impacts across the three different scopes.

### 4.1 Criteria Air Pollutants at the Source

We first model a computing task's criteria air pollutants at the source across the three different scopes in Section 3.

#### 4.1.1 Scope 1.

Onsite diesel generators are sized based on the data center power capacity, routinely tested to ensure a high availability of the entire data center, and used for demand response. Thus, the overall scope-1 air pollutants should be attributed to each computing task based on its power allocation and duration. Suppose that the overall scope-1 emission by an AI data center under consideration is  $\bar{p}^s = (\bar{p}_1^s, \dots, \bar{p}_M^s)$ , for  $M$  types of air pollutants, over a timespan of  $\bar{T}$  (e.g., one year). Considering a task that is allocated a fraction of  $x \in (0, 1]$  of the overall data center power capacity and lasts for a duration of  $T$ , we express the scope-1 air pollutants attributed to the task as

$$p^s = \frac{x \cdot T}{\bar{T}} \cdot \bar{p}^s, \quad (1)$$

which attributes the overall emission  $\bar{p}^s$  to the task in proportion to its allocated power and duration.

#### 4.1.2 Scope 2.

A computing task's scope-2 air pollutants come from its usage of electricity generated from fossil fuels. Suppose that the power grid serving the data center has an emission rate of  $\gamma = (\gamma_1, \dots, \gamma_M)$  for  $M$  types of air pollutants to produce each unit of electricity. In practice, the power grid consists of multiple interconnected power plants to supply electricity to many customers over a wide area (e.g., a balancing area [76]). Thus, similar to carbon footprint accounting [77], the air pollutant emission rate  $\gamma$  can be calculated based on either the weighted average emission rate of all the power plants (i.e.,  $\gamma = \frac{\sum_k \gamma_k \cdot b_k}{\sum_k b_k}$  where  $\gamma_k$  and  $b_k$  are

the emission rate and electricity generation of the power plant  $k$ ) or the emission rate of the marginal power plant (i.e., the power plant dispatched in response to the next electricity demand increment), which are referred to as average emission rate or marginal emission rate, respectively. The average emission represents a proportional share of the overall air pollutant emission by an electricity consumer, while the marginal emission is useful for quantifying the additionality of air pollutants due to a consumer's electricity usage. Suppose that the electricity consumption by the computing task is  $e$ , including the data center overhead captured by the power usage effectiveness. Then, we write the scope-2 air pollutants as

$$p^s = e \cdot \gamma, \quad (2)$$

which is either based on either average or marginal accounting.

To evaluate the public health impacts of U.S. data centers, we consider the average attribution method unless otherwise noted, which is also the standard methodology of carbon emission accounting used by technology companies in their sustainability reports [10, 37, 78].

We can also refine the calculation of scope-2 air pollutants in (2) by considering the summation of air pollutants over multiple time slots over the task's duration.

**Location-based emission.** There are two types of scope-2 carbon emission accounting associated with electricity consumption: location-based and market-based [10]. Specifically, location-based carbon emissions refer to the physical carbon emissions attributed to an electricity consumer connected to the power grid, while market-based carbon emissions are net emissions after applying reductions due to contractual arrangements and other credits (e.g., renewable energy credits). As noted by a recent study on carbon accounting [79], location-based accounting is considered *essential*, whereas market-based accounting is *valuable*. Moreover, market-based accounting relies on market instruments whose detailed information is often not publicly disclosed. Thus, in this paper, we follow the literature [8] and focus on location-based accounting for scope-2 criteria air pollutants without considering market-based pollution reduction mechanisms.

We also note that market-based emission reduction is likely less effective to mitigate the public health impact. The reason is that, unlike carbon emissions that have a similar effect on climate change regardless of the emission locations, the public health impact of criteria air pollutants heavily depends on the location of the emission source. For example, the public health impact of using pollutant-intensive electricity generated from a populated region may not be effectively mitigated by the clean energy credits generated elsewhere.

#### 4.1.3 Scope 3.

While our empirical analysis focuses on scope-1 and scope-2 health impacts, we present the scope-3 pollutant attribution method to provide a more complete view. Specifically, following the attribution method for scope-3 carbon emission and water consumption [9, 13], we attribute the computing hardware's air pollutants during the manufacturing process to a specific task based on the task duration. Specifically, let the hardware's expected lifespan be  $\bar{T}_0$  and the task lasts a duration of  $T$ . Considering that the  $M$  types of air pollutants for manufacturing the hardware are  $\bar{p}_0^s = (\bar{p}_{0,1}^s, \dots, \bar{p}_{0,M}^s)$  and excluding other miscellaneous pollutants (e.g., transportation), we obtain the task's scope-3 air pollutants as

$$p^s = \frac{T}{\bar{T}_0} \cdot \bar{p}_0^s. \quad (3)$$

As a server cluster includes multiple hardware components (e.g., GPU and CPU) manufactured in different locations, we apply (3) to estimate the scope-3 air pollutants for each component manufactured in a different location.

## 4.2 Air Quality Dispersion Modeling

Once emitted from their sources, criteria air pollutants can travel long distances, impacting multiple states along their paths. Unlike carbon emissions that have a similar effect on climate change regardless of the emission source locations, the public health impact of criteria air pollutants heavily depends on the location of the emission source. Generally, the closer a receptor is to the source, the greater the air quality impact. Furthermore, the dispersion of air pollutants is influenced by meteorological conditions, such as wind speed and direction.

In practice, dispersion modeling tools are used to track the movement of air pollutants. These tools employ complex mathematical equations to simulate the atmospheric processes governing the dispersion. By

incorporating emission data and meteorological inputs, dispersion modeling can predict pollutant concentrations at selected receptor locations [80]. We consider a general dispersion modeling tool  $(p_1^r, \dots, p_N^r) = D_\theta(p^s)$ , which yields the amount of  $M$  types of air pollutants  $p_i^r = (p_{i,1}^r, \dots, p_{i,M}^r)$  at the receptor region  $i = 1, \dots, N$ . The parameter  $\theta$  captures the geographical conditions, emission source characteristics (e.g., height), and meteorological data [81]. We apply the dispersion model to each scope of air pollutants (Section 4.1) to estimate the corresponding pollutant concentrations at receptor regions.

Several dispersion modeling tools are available, including AERMOD, PCAPS and InMAP with a reduced complexity [22, 80, 82, 83]. For example, PCAPS (Pattern Constructed Air Pollution Surfaces), an advanced reduced-complexity model that provides representations of both primarily emitted PM<sub>2.5</sub> and secondarily formed PM<sub>2.5</sub> and ozone, is used in COBRA as a quick assessment of otherwise lengthy iterations and simulations of various pollution scenarios in terms of the annual average PM<sub>2.5</sub> and seasonal average maximum daily average 8-hour ozone [22, 83]. Even compared with state-of-the-science photochemical grid models, PCAPS provides similar prediction accuracies and can realistically capture the change in air pollution due to changing emissions [83]. More specifically, for electric power sectors and on-road/highway vehicle sectors (the two sectors we consider in Section 5), the prediction results of PCAPS compare very well with photochemical model predictions, with Pearson correlation coefficients of 0.92 and 0.94, respectively [22, 83].

### 4.3 Converting Health Outcomes to Economic Costs

By assessing pollutant levels  $p_i^r = (p_{i,1}^r, \dots, p_{i,M}^r)$  and population size at each receptor region  $i$ , we can estimate the incidences of health outcomes  $h_i = (h_{i,1}, \dots, h_{i,H})$  and the corresponding public health cost  $c_i = (c_{i,1}, \dots, c_{i,H})$ . The relations between  $p_i^r$  and  $h_i$  and between  $h_i$  and  $c_i$  is captured by an exposure-response function and can be established based on epidemiology research [22]. For example, the premature mortality rate can be modeled as a log-linear function in terms of the PM<sub>2.5</sub> level [84].

Further, by summing up the economic costs, we obtain quantitative estimates of the public health burden at both regional and national levels. It is important to note that the public health cost is not necessarily an out-of-pocket expense incurred by each individual, but rather reflects the estimated economic burden on a population to mitigate the adverse effects of pollutants within a specific region. Therefore, it is a quantitative scalar measure of the public health impact resulting from a particular pollutant-producing activity.

### 4.4 End-to-End Modeling

Following the end-to-end process shown in Fig. 1, we now briefly describe our modeling methodology to study the public health impact of U.S. data centers and AI training. The details are available in Appendix A.

To quantify data centers' scope-1 and scope-2 air pollutant emissions, we use air quality permit data for onsite generators [60] and electricity consumption data, including both historical records and 2028 projections, from the Lawrence Berkeley National Laboratory report [4].

To model the air pollutant dispersion and quantify health impacts, we use the latest COBRA (Desktop v5.1, as of October 2024) provided by the U.S. EPA [39]. COBRA integrates reduced-complexity air dispersion modeling (including both primarily emitted PM<sub>2.5</sub> and secondarily formed PM<sub>2.5</sub> and ozone [83]) with various concentration-response functions [22], offering a quantitative screening analysis particularly suitable for large-scale health impacts. The same or similar reduced-complexity modeling tools have been commonly used in the literature to examine the health impacts of various industries over a large area [82, 85], including electric vehicles [86], bitcoin mining [87], and inter-region electricity imports [88], among others. While each health impact model used by COBRA considers 95% confidence intervals, the high-end and low-end estimates provided by COBRA are based on different models instead of the 95% confidence interval of a single model [22].

## 5 Results

We now present our estimates of the public health impacts caused by the U.S. data centers in aggregate and by training a large generative AI model at specific locations. We focus on the contiguous U.S., which hosts nearly half of the world's data centers [38], and simply refer to it as the U.S. For consistency with COBRA, cities considered county-equivalents for census purposes are also referred to as "counties" in our paper. All our monetary values are for one year (or one computing task if applicable) and in 2023 U.S. dollars as used by COBRA.

Our results demonstrate that in 2028, the total scope-1 and scope-2 pollutants of U.S. data centers alone could cause, among others, approximately 600,000 asthma symptom cases and 1,300 premature deaths, exceeding 1/3 of asthma deaths in the U.S. each year [40]. The overall public health costs could reach more than \$20 billion, rival or even top those of on-road emissions of the largest U.S. states such as California with ~35 million registered vehicles [41]. Importantly, the health costs are unevenly distributed across counties and communities, particularly affecting low-income counties that could experience approximately 200x per-household health costs than others. Moreover, depending on the locations, training an AI model of the Llama-3.1 scale can produce an amount of air pollutants equivalent to driving a passenger car for more than 10,000 LA-NYC round trips, resulting in a health cost that even exceeds 120% of the training electricity cost.

## 5.1 Public Health Impact of U.S. Data Centers

We now present our public health impact analysis for U.S. data centers in aggregate, beginning with the historical analysis from 2019 to 2023 and highlighting the uneven distribution of health impacts, followed by a 2028 projection. The overall trend is shown in Fig. 4, which demonstrates a significant increase in the public health impact of U.S. data centers from 2023 to 2028. Specifically, the surging demand for AI data centers in the U.S. has outweighed the power plant emission efficiency improvement, potentially tripling the public health cost from 2023 to 2028.

We present the details in Table 1 and see that scope-2 health cost dominates the scope-1 cost.<sup>6</sup> This suggests that while using alternative fuels for onsite generators can help improve local air quality around data centers, greater health benefits can be achieved by powering data centers with pollutant-free electricity sources.

Mobile sources, including vehicles, marine engines, and generators, collectively account for more than half of the air pollutants in the U.S., with vehicles being a primary contributor [89,90]. Thus, to contextualize the public health costs of data centers, we compare them to on-road emissions in the three largest U.S. states (California, Texas, and Florida). In particular, California, which has about 35 million registered vehicles, exhibits the highest public health cost from on-road emissions among all U.S. states [39,41]. In the COBRA model, on-road emissions are categorized under the “Highway Vehicles” sector and include both tailpipe exhaust and tire and brake wear. The details of calculating on-road emissions and the corresponding health costs are provided in Appendix A.

As shown in Fig. 4, in 2023, the total public health cost of U.S. data centers is 42% of that from California’s on-road emissions. Due to the tightening air pollutant regulations [91], the health costs of on-road emissions have generally decreased from 2019 to 2028. However, with rapid growth, the public health impact of U.S. data centers is projected to rival or even surpass that of California’s on-road emissions by 2028, underscoring the need for greater attention to the growing public health impacts of the data center industry.

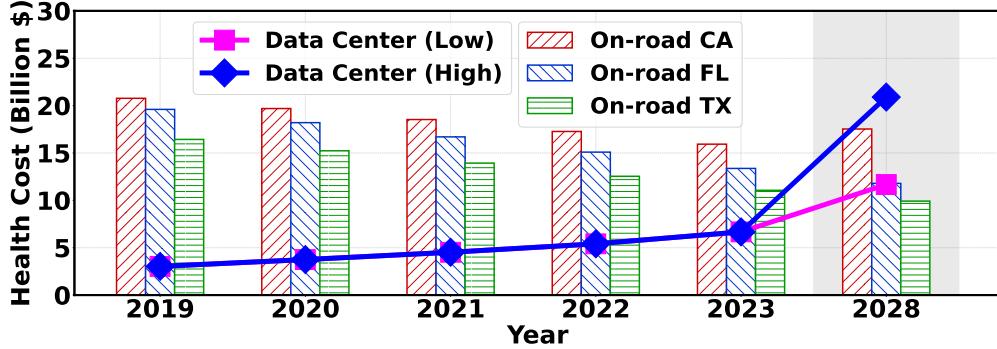
### 5.1.1 Historical analysis: 2019-2023.

Table 1 shows the public health cost of U.S. data centers from 2019 to 2023 as a reference. Even at the beginning of the generative AI boom, the U.S. data centers have already resulted in a total public health cost of about \$6.7 billion, or \$47.5 per household, in 2023. This is equivalent to approximately 44% of the data centers’ total electricity cost.

Next, we show in Fig. 5 the county-level total public health cost of U.S. data centers from 2019 to 2023, which exhibits significant spatial variability. In particular, populated counties located downwind of power plants supplying electricity to data centers tend to experience higher health costs, reflecting the transport of air pollutants across regions. The cumulative distribution function (CDF) in Fig. 5b highlights that while many counties incur relatively low health costs, a small fraction of counties bear substantially higher impacts. Table 5c further identifies the top-10 counties with the highest total health costs, illustrating how local population density and proximity to power generation infrastructure combine to amplify public health risks in specific communities.

---

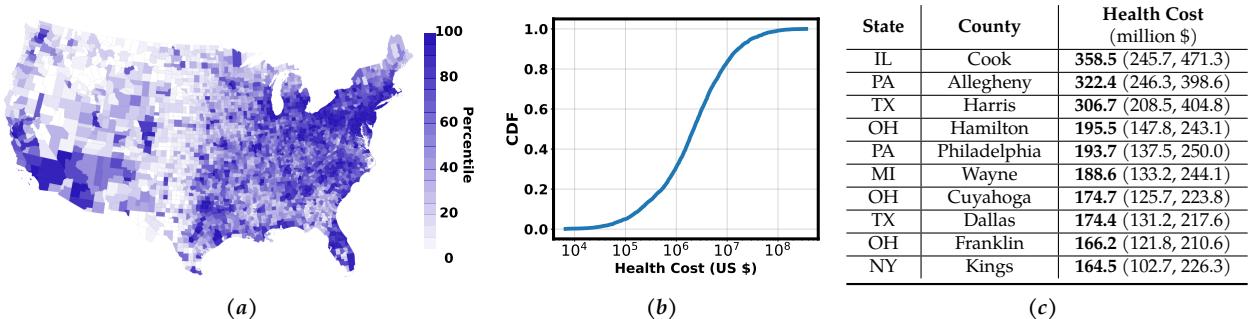
<sup>6</sup>We use the “mid (low, high)” format to represent the midrange, low and high estimates offered by COBRA. When presenting a single value or a ratio (e.g., health-to-electricity cost ratio), we use the midrange by default.



**Figure 4:** The public health costs of U.S. data centers and top-3 state on-road emissions from 2019 to 2023 and the 2028 projection based on the Lawrence Berkeley National Lab’s report [4]. The cost for U.S. data centers includes scope-1 and scope-2 impacts. The “High” and “Low” represent the high and low growth rates considered in [4].

**Table 1:** The public health cost of U.S. data centers from 2019 to 2023 and projection in 2028

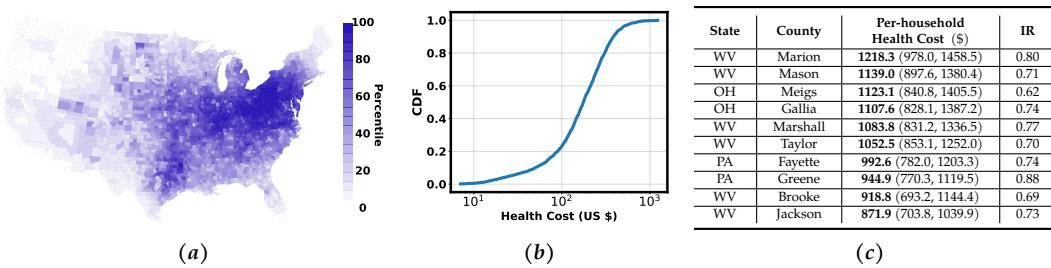
Year	Electricity (TWh)	Electricity Cost (billion \$)	Scope	Mortality	Health Cost (billion \$)	Per-Household Health Cost (\$)	% of CA On-road Health Cost
2019	90.28	7.73	Scope-1	7 (6, 8)	0.11 (0.10, 0.13)	0.84 (0.72, 0.97)	1%
			Scope-2	189 (138, 240)	2.92 (2.18, 3.66)	21.51 (16.05, 26.97)	14%
			Total	196 (144, 248)	3.03 (2.27, 3.79)	22.36 (16.77, 27.94)	15%
2020	105.56	9.04	Scope-1	9 (8, 11)	0.15 (0.13, 0.17)	1.11 (0.94, 1.27)	1%
			Scope-2	233 (171, 296)	3.60 (2.68, 4.51)	26.30 (19.62, 32.98)	18%
			Total	243 (179, 307)	3.75 (2.81, 4.69)	27.41 (20.56, 34.25)	19%
2021	127.78	10.94	Scope-1	13 (11, 15)	0.20 (0.17, 0.24)	1.48 (1.26, 1.70)	1%
			Scope-2	280 (205, 355)	4.31 (3.22, 5.41)	31.24 (23.31, 39.17)	23%
			Total	293 (216, 370)	4.52 (3.39, 5.64)	32.72 (24.57, 40.88)	24%
2022	151.39	12.97	Scope-1	21 (17, 24)	0.33 (0.28, 0.38)	2.40 (2.04, 2.76)	2%
			Scope-2	330 (242, 418)	5.08 (3.79, 6.37)	36.44 (27.18, 45.70)	29%
			Total	351 (259, 443)	5.41 (4.07, 6.75)	38.84 (29.22, 48.46)	31%
2023	176.39	15.11	Scope-1	32 (26, 37)	0.51 (0.43, 0.59)	3.65 (3.08, 4.21)	3%
			Scope-2	401 (294, 508)	6.16 (4.59, 7.73)	43.83 (32.69, 54.97)	39%
			Total	433 (320, 546)	6.67 (5.03, 8.32)	47.48 (35.77, 59.19)	42%
2019 to 2023	651.40	55.79	Scope-1	82 (68, 95)	1.32 (1.12, 1.52)	9.49 (8.05, 10.92)	1%
			Scope-2	1434 (1050, 1818)	22.07 (16.46, 27.67)	159.32 (118.85, 199.80)	24%
			Total	1516 (1118, 1913)	23.38 (17.58, 29.19)	168.81 (126.90, 210.72)	25%
2028 (Low)	325.00	27.84	Scope-1	54 (46, 63)	0.90 (0.77, 1.03)	6.11 (5.22, 7.00)	5%
			Scope-2	650 (483, 818)	10.78 (8.14, 13.41)	73.29 (55.37, 91.21)	61%
			Total	705 (529, 880)	11.67 (8.91, 14.44)	79.40 (60.59, 98.21)	67%
2028 (High)	580.00	49.68	Scope-1	97 (82, 112)	1.61 (1.37, 1.84)	10.94 (9.35, 12.53)	9%
			Scope-2	1165 (865, 1464)	19.29 (14.58, 24.01)	131.23 (99.14, 163.31)	110%
			Total	1262 (947, 1576)	20.90 (15.95, 25.85)	142.16 (108.49, 175.84)	119%



**Figure 5:** The county-level total health cost of U.S. data centers from 2019 to 2023. (a) Health cost map; (b) CDF of county-level health cost; (c) Top-10 counties by total health cost.

### 5.1.2 Uneven distribution of data centers’ public health impacts.

Next, Fig. 6 presents the county-level per-household total health costs attributable to U.S. data centers from 2019 to 2023. The results reveal a highly disproportionate distribution of health impacts across counties, with low-income communities particularly affected. The ratio of the highest to lowest county-level per-household health cost reaches approximately 200. Notably, all of the top 10 counties with the highest per-household costs have median household incomes below the national median. The high degree of disparity across



**Figure 6:** The county-level per-household health cost of U.S. data centers from 2019 to 2023. (a) Per-household health cost map; (b) CDF of county-level per-household health cost; (c) Top-10 counties by per-household health cost. IR represents “County-to-Nation Per-Household Median Income Ratio.”

different communities in terms of the public health cost suggests that we need to carefully examine the local and regional health impacts of data centers and improve public health equity to enable truly responsible computing.

Furthermore, many of the hardest-hit communities neither host large data centers nor directly benefit economically from AI data centers, such as through tax revenues. For example, several counties in West Virginia are among the most affected, because many coal-fired power plants in West Virginia are supplying electricity to data centers in the neighboring state of Virginia [4, 5]. By contrast, despite hosting a large number of data centers and incurring high total health costs due to its large population, California’s clean power grid results in some of the lowest per-household health impacts in the country.

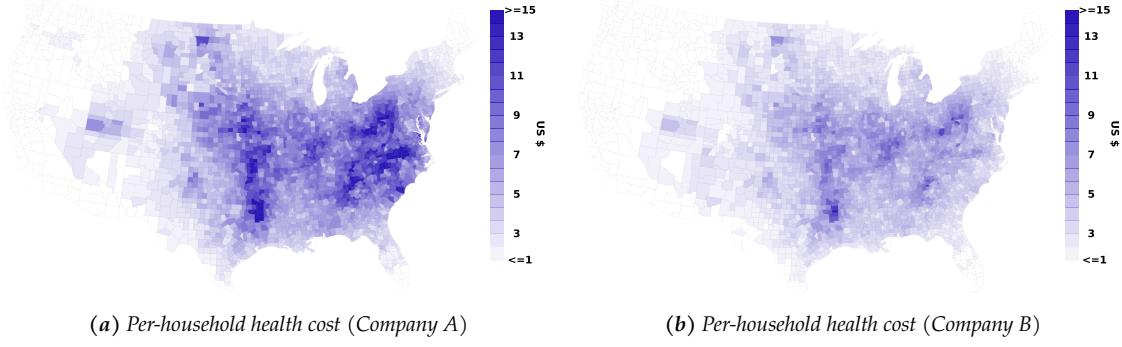
### 5.1.3 Projection for 2028.

According to a recent Lawrence Berkeley National Laboratory (LBNL) report [4], the U.S. data center electricity consumption is expected to increase from 4.4% of the total national electricity use in 2023 to 6.7–12.0% in 2028, depending on the growth trajectory of AI adoption. At the same time, the projected rise in peak power demand is accompanied by massive installations of onsite backup diesel generators to ensure reliability during grid contingencies [60].

This substantial growth in electricity demand and onsite generation is expected to offset, and in fact outweigh, the gradual pollution emission intensity reductions anticipated from the power sector. As a result, the total public health costs attributable to data center operations are projected to potentially triple from 2023 to 2028. Quantitatively, based on the low- and high-growth scenarios considered in [4], the total public health impact of U.S. data centers is estimated to reach \$11.7 billion and \$20.9 billion in 2028, respectively. Under the high-growth scenario, the resulting health burden could rival or exceed that of on-road emissions in the largest U.S. state. This highlights the growing health externalities of U.S. data centers at a national scale.

**Table 2:** The public health cost of training a large AI model in selected U.S. data centers.

Location	Electricity Price (¢/kWh)	Electricity (million \$)	Health Cost (million \$)	% of Electricity Cost	Emission (Metric Ton)		
					PM2.5 (LA-NYC)	NOx (LA-NYC)	SO2
Huntsville, AL	7.11	2.1	0.70 (0.54, 0.87)	33%	0.61 (13800)	2.80 (2500)	2.72
Stanton Springs, GA	6.88	2.0	0.85 (0.65, 1.04)	41%	0.69 (15500)	3.37 (3000)	3.35
DeKalb, IL	8.20	2.4	1.92 (1.41, 2.42)	79%	1.25 (28100)	7.31 (6600)	7.83
Altoona, IA	6.91	2.1	2.51 (1.84, 3.17)	122%	1.52 (34000)	11.78 (10600)	14.76
Sarpy, NE	7.63	2.3	1.54 (1.16, 1.92)	68%	1.13 (25300)	13.5 (12200)	18.51
Los Lunas, NM	5.75	1.7	0.73 (0.56, 0.90)	43%	0.78 (17500)	8.36 (7500)	9.84
Forest City, NC	7.15	2.1	1.07 (0.85, 1.30)	50%	0.72 (16200)	5.72 (5200)	3.27
New Albany, OH	7.03	2.1	1.61 (1.20, 2.03)	77%	1.13 (25200)	5.15 (4600)	4.44
Prineville, OR	7.52	2.2	0.23 (0.19, 0.28)	10%	0.59 (13300)	4.67 (4200)	2.40
Gallatin, TN	6.23	1.9	0.32 (0.24, 0.40)	17%	0.41 (9200)	1.21 (1100)	0.93
Fort Worth, TX	6.60	2.0	0.51 (0.38, 0.65)	26%	0.47 (10500)	3.02 (2700)	3.81
Eagle Mountain, UT	6.99	2.1	0.24 (0.19, 0.29)	12%	0.60 (13300)	4.82 (4300)	2.52
Henrico, VA	8.92	2.7	1.61 (1.20, 2.03)	61%	1.13 (25200)	5.15 (4600)	4.44



*Figure 7: The county-level per-household health cost of two U.S. technology companies in 2023.*

## 5.2 Public Health Impact of Generative AI Training

We now study the health impact of a specific computing task. Specifically, we consider the training of an LLM and assume that the electricity consumption is the same as training Llama-3.1 recently released by Meta [92]. As the scope-2 impact is dominant and the power allocated to train Llama-3.1 is unknown to determine scope-1 impacts, we focus on scope-2 health costs associated with the electricity consumption. While we use Meta’s Llama-3.1 training electricity consumption and U.S. data center locations as an example, our results should be interpreted as the estimated public health impact of training a general LLM with a comparable scale of Llama-3.1.

We show the results in Table 2. It can be seen that the total health cost can even exceed 120% of the electricity cost and vary widely depending on the training data center locations. For example, the total health cost is only \$0.23 million in Oregon, whereas the cost will increase dramatically to \$2.5 million in Iowa due to various factors, such as the wind direction and the pollutant emission rate for electricity generation [76]. Additionally, depending on the locations, training an AI model of the Llama-3.1 scale can produce an amount of air pollutants equivalent to more than 10,000 LA-NYC round trips by car.

The results highlight that the public health impact of AI model training is highly location-dependent. Combined with the spatial flexibility of model training, they suggest that AI model developers should take into account potential health impacts when choosing data center locations for training.

## 5.3 Location-Dependent Public Health Impacts of Two Technology Companies

We further highlight locational dependency of public health impacts by considering two major technology companies’ U.S. data center locations in 2023, excluding their leased colocation data centers whose locations are proprietary. We name these two companies A and B, respectively. These two companies do not have same data center locations. While company B discloses its per-location electricity usage [37], company A does not. Thus, we uniformly distribute company B’s North America electricity consumption over its U.S. data center locations based on its latest sustainability report [10]. We consider location-based emission accounting without taking into account renewable energy credits these two companies apply to offset their grid electricity consumption (see “Location-based emission” in Section 4.1.2).

We see from Fig. 7 that the two companies have significant differences in terms of the per-household health cost distribution and most-affected counties. This is primarily due to the two companies’ different data center locations, and highlights the locational dependency of public health impacts. That is, unlike carbon emissions that have a similar effect regardless of the emission source locations, the public health impact of criteria air pollutants heavily depends on the location of the emission source. Thus, technology companies should account for public health impacts when deciding where they build data centers, where they get electricity for their data centers, and where they install onsite renewables in order to best mitigate the adverse health effects.

## 6 Health-Informed Computing: Addressing Data Centers' Public Health Impact

In this section, we present Health-Informed Computing, a framework that explicitly incorporates public health impacts as a key optimization objective and strategically manages data center workloads to minimize adverse health outcomes while supporting broader sustainability goals.

To mitigate the public health impact of computing, one straightforward approach is to focus solely on reducing the energy consumption (e.g., reducing AI model sizes [93,94]). While reducing energy consumption is beneficial, overlooking the downstream public health impact of *where* and *when* energy is produced does not necessarily lead to minimized health burdens. For example, Table 2 demonstrates a 10x difference in health costs for training the same AI model across different locations. This highlights that health-informed and energy-aware computing, when combined, offer complementary benefits, leading to better public health outcomes.

### 6.1 Opportunities for Health-Informed Computing

Data centers, including those operated by major technology companies [10,37], predominantly rely on grid electricity due to the practical challenges of installing on-site low-pollutant and low-carbon energy sources at scale. However, the spatial-temporal variations of scope-2 health costs (Fig. 8) open up new opportunities to reduce the public health impact by exploiting the high scheduling flexibilities of computing workloads (e.g., AI training). For example, as further supported by EPRI's recent initiative on maximizing data center flexibility for demand response [11], AI training can be scheduled in more than one data center, while multiple AI models with different sizes are often available to serve AI inference requests, offering flexible resource-performance tradeoffs.

To date, the existing data centers have mostly exploited such scheduling flexibilities for reducing electricity costs [95], carbon emissions [15], water consumption [96], and/or environmental inequity [97]. Nonetheless, the public health impact of AI significantly differs from these environmental costs or metrics.

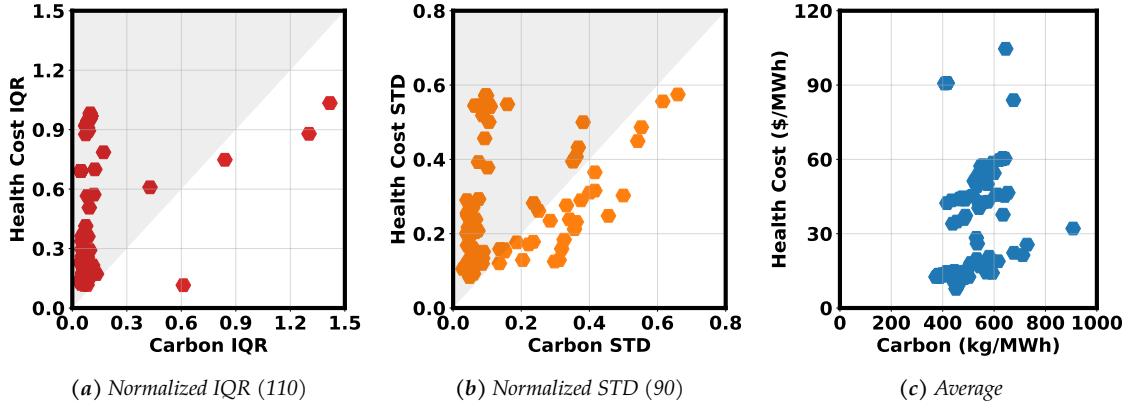
Concretely, despite sharing some common sources (e.g., fossil fuels) with carbon emissions, the public health impact resulting from the dispersion of criteria air pollutants is highly dependent on the emission source location and only exhibits a weak correlation with carbon emissions. For example, the same quantity of carbon emissions generally results in the same climate change impacts regardless of the emission source; in contrast, criteria air pollutants have substantially greater public health impacts if emitted in densely populated regions compared to sparsely populated or unpopulated regions, emphasizing the importance of considering spatial variability.

To further confirm this point and highlight the potential of health-informed data center load shifting, we analyze the scope-2 marginal carbon intensity and public health cost for each unit of electricity generation across all the 114 U.S. regions between October 1, 2023, and September 30, 2024, provided by WattTime [77].<sup>7</sup> The time granularity for data collection is 5 minutes.

Here, we focus on marginal health impacts and carbon emissions for two main reasons: first, WattTime provides only real-time marginal health impact estimates [98]; and second, marginal signals are generally considered more useful for guiding energy load adjustments [99], which also explain why the EPA reports marginal health benefits per kWh (i.e., marginal health price) to inform energy demand changes [100].

We show in Fig. 8a the region-wise normalized interquartile ranges (IQR divided by the yearly average) for both public health costs and carbon emissions. The normalized IQR measures the spread of the time-varying health and carbon signals. Specifically, in 110 out of the 114 U.S. regions (96%), the normalized IQR of health cost is higher than that of the carbon intensity for each unit of electricity consumption. Moreover, the normalized IQR for carbon emissions is less than 0.2 in most of the regions. This implies that health costs exhibit a greater temporal variation than carbon emissions in 110 out of the 114 U.S. regions. Likewise, in Fig. 8b, the greater temporal variation of health costs is also supported by its greater normalized standard deviation (STD divided by the yearly average) in 90 out of the 114 U.S. regions (79%). Next, we show in Fig. 8c the weak spatial correlation (Pearson correlation coefficient: 0.292) between the yearly average health cost and carbon intensity across the 114 regions. Furthermore, the normalized IQR of the health cost spatial

<sup>7</sup>The health cost signal provided by [77] only considers mortality from PM<sub>2.5</sub>, while COBRA includes a variety of health outcomes including asthma, lung cancer, and mortality from ozone, among others [22].



**Figure 8:** Analysis of marginal scope-2 carbon emission rates and public health costs over 114 U.S. regions between October 1, 2023 and September 30, 2024 [77]. (a) In 110 out of the 114 U.S. regions (96%), the normalized IQR of marginal health cost is higher than that of marginal carbon intensity. (b) In 90 out of the 114 U.S. regions (79%), the normalized standard deviation of marginal health cost is higher than that of marginal carbon intensity. (c) The Pearson correlation between the per-region yearly average marginal health cost and carbon intensity is 0.292.

Location	Pearson Correlation	Normalized IQR			Normalized STD		
		Health	Carbon	Health:Carbon Ratio	Health	Carbon	Health:Carbon Ratio
Loudoun County, VA	0.427	0.158	0.065	2.409	0.131	0.059	2.222
Central Ohio, OH	0.479	0.160	0.065	2.441	0.137	0.066	2.064
The Dalles, OR	0.326	0.957	0.099	9.614	0.546	0.103	5.296
Douglas County, GA	0.756	0.507	0.093	5.418	0.293	0.075	3.913
Montgomery County, TN	0.760	0.289	0.067	4.320	0.195	0.046	4.236
Papillion, NE	0.736	0.748	0.840	0.891	0.487	0.553	0.881
Storey County, NV	0.584	0.178	0.057	3.132	0.168	0.042	4.004
Ellis County, TX	0.474	0.196	0.082	2.384	0.232	0.361	0.641
Berkeley County, SC	0.416	0.156	0.054	2.911	0.105	0.044	2.405
Council Bluffs, IA	0.361	0.185	0.111	1.671	0.129	0.311	0.415
Henderson, NV	0.584	0.178	0.057	3.132	0.168	0.042	4.004
Jackson County, AL	0.760	0.289	0.067	4.320	0.195	0.046	4.236
Lenoir, NC	0.240	0.176	0.059	2.982	0.129	0.046	2.800
Mayes County, OK	0.617	0.122	0.049	2.495	0.171	0.222	0.772

**Table 3:** Correlation analysis of marginal carbon emissions and health impacts for a technology company’s U.S. data center locations between October 1, 2023, and September 30, 2024 [77]. According to the region classification of WattTime [98], the two data centers in Storey County, NV, and Henderson, NV, belong to the same power grid region, and so do those in Jackson County, AL, and Montgomery County, TN.

distribution is 3.62x that of carbon emission spatial distribution (1.05 vs. 0.29), while the health-to-carbon ratio in terms of the spatial distribution’s normalized STD is 3.37 (0.64 vs. 0.19). In other words, the health cost could have a greater spatial spread than the carbon emission.

In addition to analysis for all U.S. regions, we turn to specific regions where a large technology company builds its U.S. data centers. We present the results Table 3, further confirming that carbon intensities and health impacts are not always aligned and that health impacts vary more significantly than carbon intensities in almost all the locations.

We further analyze the Pearson correlation coefficients between hourly marginal health prices and carbon emission rates throughout 2023 for the U.S. regions that have complete health and carbon data provided by WattTime [77]. The CDF of the correlation coefficients is shown in Figure 9a. We see that nearly 70% of the regions have a weak or moderate correlation, with a carbon-health correlation coefficient of less than 0.60. This implies that despite having fossil fuels as the common source, health costs and carbon emissions are different and can exhibit trade-offs. Moreover, due to their additional dependence on population distribution and meteorological conditions, health prices demonstrate more pronounced temporal fluctuations than carbon emissions.

These findings highlight that leveraging spatiotemporal variations in a health-aware manner can substantially reduce the public health costs of data center operations. Moreover, the observed distinctions between health impacts and carbon emissions suggest the need to optimize data center decisions by explicitly accounting for and exploiting the spatiotemporal heterogeneity of health impacts.

## 6.2 Benefits of Health-Informed Computing

To improve system performance and reliability, technology companies typically operate data centers over a variety of geographically distributed regions and *dynamically* distribute computing workloads through a process known as geographical load balancing (GLB). Here, we leverage the unique spatial load flexibility of geographically distributed data centers to demonstrate the benefits of health-informed computing as a proof of concept.

**Health-Informed GLB.** Specifically, we study health-informed GLB (HI-GLB) to address the public health burden of data center operation. We consider a discrete-time model of duration  $T$  and measure the workloads in terms of their energy demand. For the sake of examining the impact of spatial flexibility, we assume that the energy loads (i.e., workloads) can be flexibly distributed across a set of  $N$  data centers denoted as  $\mathcal{N} = \{1, 2, \dots, N\}$ . In each time  $t \in \{1, 2, \dots, T\}$ , the total energy demand is  $M_t$ , and  $w_{i,t}$  represents the load assigned to data center  $i$ . We use  $l_i$  to represent the default load capacity of data center  $i$ , and introduce a slackness parameter  $\lambda \geq 1$  to represent the ability of each data center to accept loads in excess of its default capacity. Thus, the greater the value of  $\lambda$ , the more spatial flexibility the operator has. In each time  $t$ , we use  $p_{i,t}^e$  and  $p_{i,t}^h$  to denote the electricity price and health price at data center  $i$ , respectively.

The health price  $p_{i,t}^h$  depends on the emission source of criteria air pollutants (e.g., power plants' emission rates and their locations), air pollutant dispersion, and estimates of adverse health outcomes and resulting costs attributed to the increased air pollutant concentration in each region [22]. Thus, the health price quantifies the ultimate health burden imposed on affected populations and is measured in terms of economic costs for each unit of electricity consumption. It varies over time due to fluctuations in the grid's generation mix and changing meteorological conditions. Third-party organizations such as WattTime [98] provide real-time estimates of the marginal health price of electricity across 114 power balancing regions in the U.S., while the EPA [100] reports annualized average health prices for electricity in 14 broader regions nationwide.

Our goal is to minimize the sum of electricity costs  $\sum_{t=1}^T \sum_{i=1}^N p_{i,t}^e \cdot w_{i,t}$  and health costs  $\sum_{t=1}^T \sum_{i=1}^N p_{i,t}^h \cdot w_{i,t}$  across all regions. Thus, HI-GLB is formulated as follows:

$$\min_w \quad \sum_{t=1}^T \sum_{i=1}^N p_{i,t}^e \cdot w_{i,t} + \sum_{t=1}^T \sum_{i=1}^N p_{i,t}^h \cdot w_{i,t} \quad (4a)$$

$$\text{s.t.} \quad \sum_{i=1}^N w_{i,t} = M_t, \quad \forall t \in \mathcal{T} \quad (4b)$$

$$0 \leq w_{i,t} \leq \lambda \cdot l_i, \quad \forall i \in \mathcal{N}, \forall t \in \mathcal{T} \quad (4c)$$

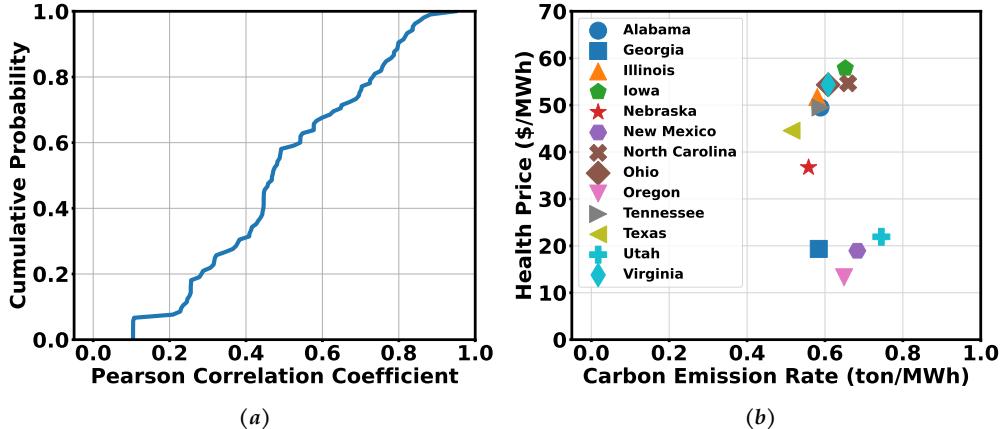
where the constraint (4b) means that all loads must be dispatched to a data center (with no loads dropped) and the constraint (4c) encodes the maximum workload capacity of each data center.

Our formulation can be easily extended to incorporate additional considerations such as load dispatching distance constraints and other metrics such as carbon emissions. Additionally, it can also include long-term, per-region health impact constraints, rather than focusing solely on national-level total health costs. For the sake of clarity, we set these extensions aside to focus on the novel metric of health cost for data center resource management.

We use Meta's electricity consumption for each U.S. data center location in 2023 [37] as the baseline. Our study is intended to illustrate the potential benefits of health-informed GLB and should not be interpreted as representing Meta's actual health impacts. For comparison, we consider carbon-aware GLB which minimizes the cost objective  $\sum_{t=1}^T \sum_{i=1}^N p_{i,t}^e \cdot w_{i,t} + \sum_{t=1}^T \sum_{i=1}^N p^c \cdot r_{i,t}^c \cdot w_{i,t}$ , where  $r_{i,t}^c$  is the carbon emission rate and  $p^c$  is the carbon price to encourage carbon reduction. We vary the carbon price  $p^c$  from \$5/ton to \$200/ton, which is consistent with the range adopted by the U.S. federal government over several previous administrations [101]. Additionally, as two special cases, we include \$0/ton and  $\$ \infty / \text{ton}$ , which represent the respective cases of purely optimizing electricity costs and purely minimizing carbon emissions. While

**Table 4:** Comparison between health-informed and carbon-aware GLB ( $\lambda = 1.5$ )

Metric	Baseline	Carbon-Aware GLB				HI-GLB ( $p_{i,t}^e = 0$ )	HI-GLB
		$p^c = \$0/\text{ton}$	$p^c = \$5/\text{ton}$	$p^c = \$200/\text{ton}$	$p^c = \$\infty/\text{ton}$		
Health (Million \$)	393.23	416.29 (5.86%)	416.29 (5.86%)	383.32 (-2.52%)	404.26 (2.80%)	289.67 (-26.34%)	291.94 (-25.76%)
Energy (Million \$)	756.50	714.99 (-5.49%)	714.99 (-5.49%)	734.77 (-2.87%)	765.68 (1.21%)	741.49 (-1.98%)	733.66 (-3.02%)
Carbon (Million Ton)	6.60	6.67 (1.02%)	6.67 (1.02%)	6.20 (-6.01%)	6.12 (-7.23%)	6.54 (-0.89%)	6.51 (-1.38%)



**Figure 9:** Correlation analysis. (a) CDF of correlation coefficients between hourly health prices and marginal carbon emission rates for all the U.S. regions; (b) Scatter plot of health price and marginal carbon emission rate (annual average in 2023) across Meta’s U.S. data center locations.

joint consideration of carbon-aware and health-informed GLB is interesting, we exclude such an analysis to better contrast the differences between existing carbon-aware GLB approaches and our proposed health-informed GLB framework.

For each data center location, we use the health price provided by WattTime [77] and the industry electricity price from the EIA [102]. To quantify the carbon emissions of each location, we consider the marginal carbon emissions rate provided by WattTime [77]. We specifically use marginal carbon emissions for a fair comparison with the marginal health price, and since this metric often better reflects the actual impact on system-wide carbon emissions resulting from demand-side load shifting [99]. More details can be found in Appendix B.1.

**Health-Informed vs. Carbon-Aware GLB.** We show results comparing GLB with health and carbon costs in Table 4. The results highlight that optimizing solely for environmental metrics like carbon does not effectively mitigate health impacts. Concretely, as the carbon price varies, the health cost ranges between \$380 million and \$420 million, whereas the baseline health cost is approximately \$393 million. When  $p_c = \$200/\text{ton}$ , the health cost decreases by 2.52% relative to the baseline. However, for all other carbon prices, the health cost increases by using carbon-aware GLB. Among these, the algorithm that focuses solely on minimizing carbon ( $p_c = \infty$ ) leads to a 2.80% increase in health cost compared to the baseline. This stands in sharp contrast to HI-GLB, which achieves a substantial  $\sim 26\%$  reduction in health cost compared to the baseline, demonstrating that environmental metrics such as carbon are insufficient for mitigating health impacts unless public health considerations are explicitly integrated into the optimization process. Moreover, compared to the pure health-informed GLB that sets a zero electricity price ( $p^e = 0$ ), HI-GLB only slightly increases the health cost from \$290 million to \$292 million, further demonstrating that considering the health price can effectively drive healthier GLB decisions while maintaining low cost.

At the same time, the pure carbon-aware GLB algorithm achieves a maximum reduction in carbon reduction of 7.23% while increasing the health burden and electricity cost relative to the baseline. In contrast, HI-GLB achieves 26% health cost reduction, 3% electricity cost saving, and more than 1% carbon reduction. This highlights the necessity of adopting health-informed algorithms for health burden reduction while offering co-benefits of cost saving and carbon reduction.

We also vary the capacity slackness  $\lambda$  to change the spatial flexibility and obtain similar insights. The full details are available in Appendix B.2.

**Spatial correlation analysis.** To further explain the results, we analyze the correlation between health costs and carbon emissions across the 13 regions where Meta operates its U.S. data centers. Figure 9b presents a scatter plot illustrating the correlation between the annual average carbon emission rate and health price across various regions. The Pearson correlation coefficient between different locations' health prices and carbon emissions is approximately -0.35, indicating a negative relationship. This suggests a potential conflict between efforts to optimize carbon emissions and those aimed at improving health outcomes via GLB. Furthermore, as highlighted by the spatial patterns in Figure 9b, the health prices exhibit a significantly higher degree of spatial variability compared to carbon emissions across regions. This discrepancy further reinforces the point that focusing solely on optimizing environmental factors, such as carbon emissions, may not effectively reduce health costs, underscoring the importance of integrating health-informed optimization strategies to achieve more comprehensive benefits.

## 7 Our Recommendations

We provide additional recommendations to address the growing public health impact of data centers.

### Recommendation 1: Standardization of Reporting Protocols

Despite their immediate and tangible impacts on public health, criteria air pollutants have been entirely overlooked in AI model cards and sustainability reports published by technology companies [10, 36, 37]. The absence of such critical information adds substantial challenges to accurately identifying specific AI data centers as a key root cause of public health burdens and could potentially pose hidden risks to public health. To enhance transparency and lay the foundation for truly responsible AI, we recommend standardization of reporting protocols for criteria air pollutants and the public health impacts across different regions. Concretely, criteria air pollutants can be categorized into three different scopes (Section 3), and reported following the greenhouse gas protocol widely adopted by technology companies [10, 37, 78].

Just as addressing scope-2 and scope-3 carbon emissions is important for mitigating climate change, it is equally crucial to address scope-2 and scope-3 criteria air pollutants to promote public health throughout the power generation and hardware manufacturing processes in support of AI. For instance, power plants are dispatched based on real-time energy demand to ensure grid stability. As a result, only focusing on regulating scope-2 air pollutants at the power plant level fails to address the root cause — electricity consumption — and overlooks the potential of demand-side solutions. In contrast, recognizing scope-2 air pollutants and their associated public health impacts enables novel opportunities for health-informed AI, which, as detailed below, taps into demand-side flexibilities to holistically reduce AI's adverse public health impacts.

### Recommendation 2: Attention to All

Counties and communities located near AI data centers or supplying electricity to them often experience most significant health burdens. Nonetheless, these health impacts can extend far beyond the immediate vicinity, affecting communities hundreds of miles away [24, 25]. For example, the health impact of backup generators in northern Virginia can affect several surrounding states (Fig. 2a) and even reach as far as Florida.

While the health impact on communities where data centers operate is increasingly recognized, there has been very little, if any, attention paid to other impacted communities that bear substantial public health burdens. This disconnect leaves those communities to shoulder the public health cost of AI silently without receiving adequate support. To fulfill their commitment to social responsibility, we recommend technology companies holistically evaluate the *cross-state* public health burden imposed by their operations on all impacted communities, when deciding where they build data centers, where they get electricity for their data centers, and where they install renewables.

Additionally, to quantify the health effects on impacted communities with greater accuracy for potential regulatory actions, we recommend further interdisciplinary research such as cross-state air quality dispersion, health economics, and health-informed computing.

### **Recommendation 3: Promoting Public Health Equity**

The public health impact of AI is highly unevenly distributed across different counties and communities, disproportionately affecting certain (often low-income) communities [31, 103]. For example, as shown in Table 6c, all the top-10 most impacted counties in the U.S. have lower median household incomes than the national median value. The ratio of the highest county-level per-household health cost to the lowest cost is approximately 200. Therefore, it is imperative to address the substantial health impact disparities across communities.

## **8 Conclusion**

In this paper, we quantify and address the overlooked public health impact of data centers. We introduce a principled methodology to model these lifecycle pollutant emissions and quantify their associated public health impacts. Our findings suggest that the total annual public health burden of U.S. data centers could exceed \$20 billion by 2028, approaching or even surpassing the impacts of on-road vehicle emissions in California. Importantly, these health costs are not evenly distributed: disadvantaged communities bear a disproportionate share, with per-household impacts potentially up to 200 times higher than in less-affected areas. This highlights the need for targeted mitigation strategies. To this end, we propose health-informed computing, a framework that explicitly incorporates public health risk as a key metric when scheduling data center workloads, enabling more informed and equitable operational decisions.

More broadly, we recommend the adoption of standardized reporting protocols for the public health costs of data centers, alongside policies that ensure attention to all impacted communities, thereby supporting responsible, sustainable, and inclusive deployment of AI infrastructure.

## **Acknowledgement**

The authors would like to thank Susan Wierman for providing comments on an initial draft of this paper.

## **References**

- [1] U.S. Centers for Disease Control and Prevention. Artificial intelligence and machine learning: Applying advanced tools for public health. <https://www.cdc.gov/surveillance/data-modernization/technologies/ai-ml.html>.
- [2] Mary Tran. U.S. Department of State DipNote: New air quality dashboard uses AI to forecast pollution levels. <https://www.state.gov/new-air-quality-dashboard-uses-ai-to-forecast-pollution-levels/>, May 2024.
- [3] Mihaela Van der Schaar, Ahmed M Alaa, Andres Floto, Alexander Gimson, Stefan Scholtes, Angela Wood, Eoin McKinney, Daniel Jarrett, Pietro Lio, and Ari Ercole. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning*, 110:1–14, 2021.
- [4] Arman Shehabi, Sarah J. Smith, Alex Hubbard, Alex Newkirk, Nuoa Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024 United States data center energy usage report. *Lawrence Berkeley National Laboratory LBNL-2001637*, December 2024.
- [5] EPRI. Powering intelligence: Analyzing artificial intelligence and data center energy consumption. *White Paper on Technology Innovation Report*, 2024. <https://www.epri.com/research/products/3002028905>.
- [6] U.S. Department of Energy. Recommendations on powering artificial intelligence and data center infrastructure, Jul. 2024.
- [7] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable AI: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813, 2022.

- [8] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. *IEEE Micro*, 42(4):37–47, jul 2022.
- [9] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. Making AI less ‘thirsty’. *Commun. ACM*, 68(7):54–61, June 2025.
- [10] Google. Environmental report. <https://sustainability.google/reports/google-2024-environmental-report/>, 2024.
- [11] EPRI. DCFlex initiative. <https://msites.epri.com/dcflex>, 2024.
- [12] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-MOE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *ICML*, 2022.
- [13] Jennifer Switzer, Gabriel Marcano, Ryan Kastner, and Pat Pannuto. Junkyard computing: Repurposing discarded smartphones to minimize carbon. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS 2023, page 400–412, New York, NY, USA, 2023. Association for Computing Machinery.
- [14] Jaylen Wang, Daniel S. Berger, Fiodar Kazhamiaka, Celine Irvene, Chaojie Zhang, Esha Choukse, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, Ricardo Bianchini, and Akshitha Sriraman. Designing cloud servers for lower carbon. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 452–470, 2024.
- [15] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, Saurav Talukdar, Eric Mullen, Kendal Smith, MariEllen Cottman, and Walfredo Cirne. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems*, 38(2):1270–1280, 2023.
- [16] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. Reducing the carbon impact of generative AI inference (today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems, HotCarbon ’23*, New York, NY, USA, 2023. Association for Computing Machinery.
- [17] Walid A. Hanafy, Qianlin Liang, Noman Bashir, Abel Souza, David Irwin, and Prashant Shenoy. Going green for less green: Optimizing the cost of reducing cloud carbon emissions. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS ’24, page 479–496, New York, NY, USA, 2024. Association for Computing Machinery.
- [18] Intel. 2024 H1 semi-annual monitoring report (Intel Ocotillo facility). <https://www.exploreintel.com/ocotillo>, 2024.
- [19] UC Davis Air Quality Research Center. Interagency monitoring of protected visual environments. <https://airquality.ucdavis.edu/improve>.
- [20] U.S. EPA. Research on health effects from air pollution. <https://www.epa.gov/air-research/research-health-effects-air-pollution>.
- [21] Giulia Grande, Jing Wu, Petter LS Ljungman, Massimo Stafoggia, Tom Bellander, and Debora Rizzuto. Long-term exposure to PM 2.5 and cognitive decline: A longitudinal population-based study. *Journal of Alzheimer’s Disease*, 80(2):591–599, 2021.
- [22] U.S. EPA. User’s manual for the co-benefits risk assessment (COBRA) screening model. <https://www.epa.gov/cobra/users-manual-co-benefits-risk-assessment-cobra-screening-model>.

- [23] Wenhua Yu, Rongbin Xu, Tingting Ye, Michael J Abramson, Lidia Morawska, Bin Jalaludin, Fay H Johnston, Sarah B Henderson, Luke D Knibbs, Geoffrey G Morgan, et al. Estimates of global mortality burden associated with short-term exposure to fine particulate matter (PM2.5). *The Lancet Planetary Health*, 8(3):e146–e155, 2024.
- [24] U.S. EPA. What is cross-state air pollution? <https://www.epa.gov/Cross-State-Air-Pollution/what-cross-state-air-pollution>.
- [25] Jian Zhang and S. Trivikrama Rao. The role of vertical mixing in the temporal evolution of ground-level ozone concentrations. *Journal of Applied Meteorology*, 38(12):1674–1691, 1999.
- [26] Health Canada. Guidance for evaluating human health impacts in environmental assessment: Air quality. <https://iaac-aeic.gc.ca/050/documents/p80054/119376E.pdf>, 2016.
- [27] World Health Organization. Air pollution is responsible for 6.7 million premature deaths every year. <https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants>.
- [28] World Health Organization. Ambient (outdoor) air pollution. [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health), 2024.
- [29] Institute for Health Metrics and Evaluation (IHME). Global burden of disease 2021: Findings from the GBD 2021 study. <https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study>, May 2024.
- [30] U.S. EPA. Human health and environmental impacts of the electric power sector. <https://www.epa.gov/power-sector/human-health-environmental-impacts-electric-power-sector>.
- [31] U.S. EPA. Power plants and neighboring communities. <https://www.epa.gov/power-sector/power-plants-and-neighboring-communities>.
- [32] Lucas Henneman, Christine Choirat, Irene Dedoussi, Francesca Dominici, Jessica Roberts, and Corwin Zigler. Mortality risk from united states coal electricity generation. *Science*, 382(6673):941–946, 2023.
- [33] European Environment Agency. The costs to health and the environment from industrial air pollution in Europe — 2024 update. <https://www.eea.europa.eu/publications/the-cost-to-health-and-the>, 2025.
- [34] Washington Department of Ecology. Diesel pollution from data centers. <https://ecology.wa.gov/air-climate/air-quality/data-centers>.
- [35] The U.S. White House. Executive order on advancing United States leadership in artificial intelligence infrastructure. <https://www.federalregister.gov/documents/2025/01/17/2025-01395/advancing-united-states-leadership-in-artificial-intelligence-infrastructure>, January 2025.
- [36] Meta Llama. Model information. [https://github.com/meta-llama/llama-models/blob/main/models/llama3\\_1/MODEL\\_CARD.md](https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md).
- [37] Meta. Sustainability report. <https://sustainability.atmeta.com/2024-sustainability-report/>, 2024.
- [38] McKinsey. Investing in the rising data center economy. *White Paper*, January 2023.
- [39] U.S. EPA. Co-benefits risk assessment health impacts screening and mapping tool (COBRA). <https://cobra.epa.gov/>.
- [40] U.S. National Center for Health Statistics. Factstats: Asthma. <https://www.cdc.gov/nchs/fastats/asthma.htm>.

- [41] California DMV. Vehicles registered by county. <https://www.dmv.ca.gov/portal/dmv-research-reports/research-development-data-dashboards/vehicles-registered-by-county>.
- [42] Virginia Joint Legislative Audit and Review Commission. Report to the Governor and the General Assembly of Virginia: Data centers in Virginia (JLARC report 158), December 2024.
- [43] Washington Department of Ecology. Health risks from diesel emissions in the Quincy area. *Air Quality Program Report (Publication 20-02-019)*, August 2020.
- [44] U.S. EPA. Use of backup generators to maintain the reliability of the electric grid. [https://www.epa.gov/system/files/documents/2025-05/rice-memo-on-duke-energy-regulatory-interpretation-04\\_17\\_25.pdf](https://www.epa.gov/system/files/documents/2025-05/rice-memo-on-duke-energy-regulatory-interpretation-04_17_25.pdf), May 2025.
- [45] Neil M. Donahue, Allen L. Robinson, and Spyros N. Pandis. Atmospheric organic particulate matter: From smoke to secondary organic aerosol. *Atmospheric Environment*, 43(1):94–106, 2009.
- [46] U.S. EPA. Benefits and costs of the Clean Air Act 1990-2020, the second prospective study. <https://www.epa.gov/clean-air-act-overview/benefits-and-costs-clean-air-act-1990-2020-second-prospective-study>, March 2011.
- [47] U.S. EPA. Summary of the Clean Air Act. <https://www.epa.gov/laws-regulations/summary-clean-air-act>.
- [48] U.S. EPA. National ambient air quality standards (NAAQS) table. <https://www.epa.gov/criteria-air-pollutants/naaqs-table>.
- [49] California Air Resources Board. Laws and regulations. <https://ww2.arb.ca.gov/resources/documents/laws-and-regulations>.
- [50] American Lung Association. State of the air. *Report*, 2024.
- [51] World Health Organization. Ambient air pollution attributable deaths. <https://www.who.int/gho/data/indicators/indicator-details/GHO/ambient-air-pollution-attributable-deaths>, 2024.
- [52] World Health Organization. WHO global air quality guidelines. <https://www.who.int/publications/i/item/9789240034228>.
- [53] U.S. EPA. Projection of counties with monitors that would not meet in 2032. <https://www.epa.gov/system/files/documents/2024-02/2024-pm-naaqs-final-2032-projections-map.pdf>, February 2024.
- [54] John P. Stevens and Supreme Court of The United States. U.S. reports: Massachusetts v. EPA, 549 U.S. 497. *The Library of Congress*, 2007.
- [55] U.S. EPA. Climate change and human health. <https://www.epa.gov/climateimpacts/climate-change-and-human-health>.
- [56] Greenhouse Gas Protocol. Standards and guidance. <https://ghgprotocol.org/>.
- [57] U.S. EPA. Controlling air pollution from stationary engines. <https://www.epa.gov/stationary-engines>.
- [58] California Air Resources Board. Fact sheet on emergency backup generators. <https://www.aqmd.gov/home/permits/emergency-generators>.
- [59] Uptime Institute. Explaining the Uptime Institute's tier classification system (April 2021 update). <https://journal.uptimeinstitute.com/explaining-upptime-institutes-tier-classification-system/>.

- [60] Virginia Department of Environmental Quality. Issued air permits for data centers. <https://www.deq.virginia.gov/permits/air/issued-air-permits-for-data-centers>.
- [61] Hanna Pampaloni. Heat wave prompts increased data center generator use; turner pushes for tier 4 upgrades. [https://www.loudounnow.com/news/heat-wave-prompts-increased-data-center-generator-use-turner-pushes-for-tier-4-upgrades/article\\_60a48bda-dc50-4d1a-8b16-399cd4340350.html](https://www.loudounnow.com/news/heat-wave-prompts-increased-data-center-generator-use-turner-pushes-for-tier-4-upgrades/article_60a48bda-dc50-4d1a-8b16-399cd4340350.html), July 2025.
- [62] Piedmont Environmental Council. Data centers, diesel generators and air quality - PEC web map. <https://www.pecva.org/uncategorized/data-centers-diesel-generators-and-air-quality-pec-web-map/>.
- [63] Government of Canada. Wet cooling towers: Guide to reporting. <https://www.canada.ca/en/environment-climate-change/services/national-pollutant-release-inventory/report/sector-specific-tools-calculate-emissions/wet-cooling-tower-particulate-guide.html>.
- [64] Anthony S. Wexler, Chris D. Wallis, Patrick Chuang, and Mason Leandro. Assessing particulate emissions from power plant cooling towers. *California Energy Commission Final Project Report (CEC-500-2023-048)*, July 2013.
- [65] Reuters. Data center reliance on fossil fuels may delay clean-energy transition. <https://www.reuters.com/technology/artificial-intelligence/how-ai-cloud-computing-may-delay-transition-clean-energy-2024-11-21/>, November 2024.
- [66] Virginia Electric and Power Company. Integrated resource plan. [https://www.dominionenergy.com/-/media/pdfs/global/company/IRP/2024-IRP-w\\_o-Appendices.pdf](https://www.dominionenergy.com/-/media/pdfs/global/company/IRP/2024-IRP-w_o-Appendices.pdf), October 2024.
- [67] U.S. Energy Information Administration. Annual energy outlook 2025. <https://www.eia.gov/outlooks/aoe>.
- [68] Hannah Ritchie and Pablo Rosado. Electricity mix. *Our World in Data*, 2024.
- [69] Google. New nuclear clean energy agreement with Kairos Power. <https://blog.google/outreach-initiatives/sustainability/google-kairos-power-nuclear-energy-agreement/>, October 2024.
- [70] CNBC. To land Meta's massive \$10 billion data center, Louisiana pulled out all the stops. will it be worth it? <https://www.cnbc.com/2025/06/25/meta-massive-data-center-louisiana-cost-jobs-energy-use.html>, June 2025.
- [71] The Associated Press. OpenAI shows off Stargate AI data center in Texas and plans 5 more elsewhere with Oracle, Softbank. <https://apnews.com/article/openai-stargate-oracle-data-center-0b3f4fa6e8d8141b4c143e3e7f41aba1>, September 2025.
- [72] Jovan Stojkovic, Chaojie Zhang, Inigo Goiri, Josep Torrellas, and Esha Choukse. DynamoLLM: Designing LLM inference clusters for performance and energy efficiency. In *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2025.
- [73] Peng Wang, Ling-Yu Zhang, Asaf Tzachor, and Wei-Qiang Chen. E-waste challenges of generative artificial intelligence. *Nature Computational Science*, 4:818–823, October 2024.
- [74] U.S. EPA. Semiconductor industry. <https://www.epa.gov/eps-partnership/semiconductor-industry>.
- [75] McKinsey. Generative AI: The next S-curve for the semiconductor industry? *White Paper*, March 2024.
- [76] U.S. EPA. Avoided emissions and generation tool (AVERT). <https://www.epa.gov/avert>.
- [77] WattTime. <https://watertime.org/>.

- [78] Microsoft. Environmental sustainability report. <https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report>, 2024.
- [79] Ian Schneider, Hui Xu, Stephan Benecke, David Patterson, Keguo Huang, Parthasarathy Ranganathan, and Cooper Elsworth. Life-cycle emissions of AI hardware: A cradle-to-grave approach and generational trends, 2025.
- [80] U.S. EPA. Air quality dispersion modeling. <https://www.epa.gov/scram/air-quality-dispersion-modeling>.
- [81] Richard T. McNider and Arastoo Pour-Bazar. Meteorological modeling relevant to mesoscale and regional air quality applications: A review. *Journal of the Air & Waste Management Association*, 70(1):2–43, 2020.
- [82] Christopher W. Tessum, Jason D. Hill, and Julian D. Marshall. InMAP: A model for air pollution interventions. *PloS ONE*, 12(4):e0176131, 2017.
- [83] Kirk R. Baker, Heather Simon, Barron Henderson, Colby Tucker, David Cooley, and Emma Zinsmeister. Source–receptor relationships between precursor emissions and O<sub>3</sub> and PM<sub>2.5</sub> air pollution impacts. *Environmental Science & Technology*, 57(39):14626–14637, 2023.
- [84] Qian Di, Yan Wang, Antonella Zanobetti, Yun Wang, Petros Koutrakis, Christine Choirat, Francesca Dominici, and Joel D. Schwartz. Air pollution and mortality in the medicare population. *New England Journal of Medicine*, 376(26):2513–2522, 2017.
- [85] U.S. EPA. Publications that cite COBRA. <https://www.epa.gov/cobra/publications-cite-cobra>.
- [86] Jean Schmitt, Marianne Hatzopoulou, Amir FN Abdul-Manan, Heather L MacLean, and I Daniel Posen. Health benefits of US light-duty vehicle electrification: Roles of fleet dynamics, clean electricity, and policy timing. *Proceedings of the National Academy of Sciences*, 121(43):e2320858121, 2024.
- [87] Gianluca Guidi, Francesca Dominici, Nat Steinsultz, Gabriel Dance, Lucas Henneman, Henry Richardson, Edgar Castro, Falco J Bargagli-Stoffi, and Scott Delaney. The environmental burden of the United States' bitcoin mining boom. <https://pubmed.ncbi.nlm.nih.gov/39502776/>, 2024.
- [88] Eleanor M. Hennessy, Jacques A. de Chalendar, Sally M. Benson, and Inês ML Azevedo. Distributional health impacts of electricity imports in the United States. *Environmental Research Letters*, 17(6):064011, 2022.
- [89] U.S. National Park Service. Where does air pollution come from? <https://www.nps.gov/subjects/air/sources.htm>.
- [90] U.S. EPA. Clean Air Act vehicle and engine enforcement case resolutions. <https://www.epa.gov/enforcement/clean-air-act-vehicle-and-engine-enforcement-case-resolutions>.
- [91] U.S. EPA. Final rule: Multi-pollutant emissions standards for model years 2027 and later light-duty and medium-duty vehicles. <https://www.epa.gov/regulations-emissions-vehicles-and-engines/final-rule-multi-pollutant-emissions-standards-model>, April 2024.
- [92] Meta. Introducing Llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>.
- [93] DeepSeek-AI. Deepseek-v3 technical report. *arXiv* 2412.19437, 2024.
- [94] Han Cai, Chuang Gan, and Song Han. Once for all: Train one network and specialize it for efficient deployment. In *ICLR*, 2019.

- [95] Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. Cutting the electric bill for internet-scale systems. In *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, SIGCOMM '09*, page 123–134, New York, NY, USA, 2009. Association for Computing Machinery.
- [96] Mohammad A. Islam, Kishwar Ahmed, Hong Xu, Nguyen H. Tran, Gang Quan, and Shaolei Ren. Exploiting spatio-temporal diversity for water saving in geo-distributed data centers. *IEEE Transactions on Cloud Computing*, 6(3):734–746, 2018.
- [97] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. Towards environmentally equitable AI via geographical load balancing. In *e-Energy*, 2024.
- [98] WattTime. Signal: Health damage. <https://watttime.org/data-science/data-signals/health-damage/>.
- [99] Joe Gorka, Noah Rhodes, and Line Roald. ElectricityEmissions.jl: A framework for the comparison of carbon intensity signals. *arXiv* 2411.06560, 2024.
- [100] U.S. EPA. Estimating the health benefits per kilowatt-hour of energy efficiency and renewable energy. <https://www.epa.gov/statelocalenergy/estimating-health-benefits-kilowatt-hour-energy-efficiency-and-renewable-energy>, 2024.
- [101] Institute for Energy Research. EPA proposes exorbitant estimate for the social cost of carbon. <https://www.instituteforenergyresearch.org/regulation/epa-proposes-exorbitant-estimate-for-the-social-cost-of-carbon/>, 2022.
- [102] U.S. EIA. Electric power plants, capacity, generation, fuel consumption, sales, prices and customers. <https://www.eia.gov/electricity/data.php>, 2023.
- [103] U.S. EPA. About the U.S. electricity system and its impact on the environment. <https://www.epa.gov/energy/about-us-electricity-system-and-its-impact-environment>.
- [104] U.S. Department of Transportation. Estimated U.S. average vehicle emissions rates per vehicle by vehicle type using gasoline and diesel. *National Transportation Statistics Table 4-43*, June 2024.
- [105] U.S. Energy Information Administration. Annual energy outlook 2023. <https://www.eia.gov/outlooks/aoe>.
- [106] U.S. EPA. Arizona nonattainment/maintenance status for each county by year for all criteria pollutants. [https://www3.epa.gov/airquality/greenbook/anayo\\_az.html](https://www3.epa.gov/airquality/greenbook/anayo_az.html), 2024.
- [107] Intel. Ocotillo campus. <https://www.exploreintel.com/ocotillo>, 2024.
- [108] Intel. 2023-24 corporate responsibility report, 2024.
- [109] U.S. Energy Information Administration (EIA). EIA open data. <https://www.eia.gov/opendata/>.

## Appendix

### A Modeling Details

We describe the evaluation methodology used for our empirical analysis. We use the latest COBRA (Desktop v5.1, as of October 2024) provided by the U.S. EPA [39] to study the public health impact of U.S. data centers in both 2019-2023 and 2028. While COBRA uses a reduced-complexity air quality dispersion model based on a source-receptor matrix for rapid evaluation, its accuracy has been validated and the same or similar model has been commonly adopted in the literature for large-area air quality and health impact analysis [82, 85, 87, 88]. We consider county-level air pollutant dispersion throughout the contiguous U.S., which is the area currently supported by COBRA [39]. Note that cities considered county-equivalents for census purposes are also referred to as “counties” in COBRA. Throughout the paper, we use “county” without further specification.

All the monetary values are presented in the 2023 U.S. dollars unless otherwise stated. We set the discount rate as 2% in COBRA as recommended by the EPA based on the U.S. Office of Management and Budget Circular No. A-4 guidance [39]. When presenting a single value or a ratio (e.g., health-to-electricity cost ratio) if applicable, we use the midrange of the low and high estimates provided by COBRA.

COBRA provides data for county-level population, health incidence, valuation, and baseline emissions for 2016, 2023, and 2028 [39]. For the data from 2019 to 2022, we use linear interpolation as recommended by the EPA’s COBRA team.

We show in Table 5 and Table 6 the total baseline emissions of air pollutants for electricity generation and on-road traffic provided by COBRA [39]. By reducing a state’s on-road emissions to zero in COBRA, we obtain the corresponding public health cost in that state.

*Table 5: U.S. electricity generation baseline emissions from 2016 to 2028*

Year	Electricity Generation Emission (Metric Ton)			
	NOx	SO2	PM2.5	VOC
2016	1100575.41	1369417.44	111604.62	30250.76
2023	711746.94	717409.25	110878.22	34311.54
2028	695495.34	733437.11	110279.40	34446.71

*Table 6: U.S. and California on-road baseline emissions from 2016 to 2028*

Year	U.S. On-road Emission (Metric Ton)				California On-road Emission (Metric Ton)			
	NOx	SO2	PM2.5	VOC	NOx	SO2	PM2.5	VOC
2016	3293579.05	25001.53	106828.36	1680342.17	202427.66	1438.07	10197.26	89087.60
2023	1588423.83	11325.07	65742.16	996965.92	98095.76	1280.27	8144.83	54141.57
2028	1130369.84	10616.37	53455.43	758508.40	86573.30	1154.27	8276.27	44586.45

On-road emissions are categorized as the “Highway Vehicles” sector in COBRA and include both tailpipe exhaust and tire and brake wear. Thus, following the EPA and U.S. Department of Transportation classification [22, 104], PM<sub>2.5</sub> resulting from road dust is not counted as emissions of highway vehicles in our study. If the PM<sub>2.5</sub> from paved road dust (categorized as “Miscellaneous → Other Fugitive Dust → Paved Roads” in COBRA) is considered, California is still projected to have the highest state-wide public health cost of on-road vehicles among all the U.S. states.

**Electricity price.** When estimating the electricity cost for data centers in 2023 and 2038, we use the state-level average price for industrial users in [102]. The projected U.S. nominal electricity price for industrial users remains nearly the same from 2023 to 2030 (24.96 \$/MMBTu in 2023 vs. 23.04 \$/MMBTu in 2030) in the baseline case per the EIA’s Energy Outlook 2023 [105]. Thus, our estimated health-to-electricity cost ratio will be even higher if we further adjust inflation.

#### A.1 Public Health Impact of Backup Generators in Virginia

Virginia has issued a total of 174 air quality permits for data center backup generators as of December 1, 2024 [60]. More than half of the data center sites are within Loudoun County. We collect a dataset of the air quality permits: permits issued before January 1, 2023, from [62], and permits issued between January

1, 2023 and December 1, 2024, from [60]. The total permitted site-level annual emission limits are approximately 13,000 tons of NO<sub>x</sub>, 1,400 tons of VOCs, 50 tons of SO<sub>2</sub>, and 600 tons of PM<sub>2.5</sub>, all in U.S. short tons. By assuming that the actual emissions are 10% of the permitted level, the data centers in Virginia could already cause approximately 14,000 asthma symptom cases and 13-19 deaths each year, among other health implications, resulting in a total annual public health burden of \$220-300 million, including \$190-260 million incurred in Virginia, West Virginia, Maryland, Pennsylvania, New York, New Jersey, Delaware, and Washington D.C., as estimated by COBRA under the “Fuel Combustion: Industrial” sector.

The scope-1 emission information for data centers in other states is not always publicly available. Thus, when estimating scope-1 emissions for data centers in another state, we apply the emission rate (tons/MWh) derived from Virginia’s data and multiply it by the energy consumption of data centers in that state. Although this approach may introduce some estimation errors, the impact is expected to be limited because scope-2 health costs are substantially more dominant.

## A.2 Data Centers’ Scope-2 Public Health Impact

The locations of emission sources depend on the power plants supplying electricity to data centers. To evaluate the public health impacts of U.S. data centers, we focus on average attribution method, which is also the standard methodology of carbon emission accounting used by technology companies in their sustainability reports [10,37,78].

We first calculate the total data center electricity consumption  $e_{DC}$  and the overall electricity consumption (including non-data center loads)  $e_{Total}$  within each electricity region. The U.S. electricity grid is divided into 14 regions following the AVoided Emissions and geneRation Tool (AVERT, the latest version v4.3 as of October 2024) provided by the EPA [76]. We use the state-level data center electricity consumption distribution for 2023 provided by EPRI [5], scale it by the U.S. total data center electricity consumption in 2019-2023 and for the 2028 projection based on data provided by [4], and then distribute state-level electricity consumption to relevant electricity regions following the state-to-region electricity apportionment used by AVERT.

We calculate the percentage  $x\% = \frac{e_{DC}}{e_{Total}}$  of the data center electricity consumption with respect to the overall electricity consumption for each electricity region. The relationship between the health impact and emission reduction in COBRA is approximately linear. Thus, we apply a reduction by  $x\%$  to the baseline emissions of all the power plants within the respective electricity region in COBRA and estimate the corresponding county-level health impacts, including health outcomes and costs.

When assessing the health impact of generative AI training, we follow the same approach, except for changing the total data center electricity consumption to the AI model training electricity consumption.

## A.3 Public Health Impact of a Semiconductor Facility

Although semiconductor manufacturing facilities are subject to air quality regulations [74], they still pose significant risks, affecting populations across large regions. Maricopa County, AZ, has been an EPA-designated non-attainment area for several years due to its failures to meet federal air quality standards [106]. The establishment of multiple semiconductor facilities in such areas could further exacerbate air quality issues.

We consider a semiconductor manufacturing facility located in Ocotillo, a neighborhood in Chandler, Arizona [107]. By averaging the rolling 12-month air pollutant emission levels listed in the recent air quality monitoring report (as of October, 2024) [18], we obtain the annual emissions as follows: 150.4 tons of NO<sub>x</sub>, 82.7 tons of VOCs, 1.1 tons of SO<sub>2</sub>, and 28.9 tons of PM<sub>2.5</sub>. By applying these on-site emissions to COBRA under the “Other Industrial Processes” sector, we obtain a total public health cost of \$14-21 million. Additionally, the total annual energy consumption by the facility is 2074.88 million kWh as of Q2, 2024 [107]. Assuming 84.2% of the energy comes from the electricity based on the company’s global average [108], we obtain the facility’s annual electricity consumption as 1746.63 million kWh. By using the average attribution method, we further obtain an estimated health cost of \$12-17 million associated with the electricity consumption. Thus, the total health cost of the facility is \$26-39 million.

By relocating the facility from Chandler, Arizona, to a planned site in Licking County, Ohio, and assuming the same emission level and electricity consumption, we can obtain the total health cost of \$94-156 million, including \$23-36 million attributed to direct on-site emissions and \$70-120 million attributed to electricity consumption.

## A.4 Energy Consumption for Training a Generative AI Model

We consider Llama-3.1 as an example generative AI model. According to the model card [36], the training process of Llama-3.1 (including 8B, 70B, and 405B) utilizes a cumulative of 39.3 million GPU hours of computation on H100-80GB hardware, and each GPU has a thermal design power of 700 watts. Considering Meta’s 2023 PUE of 1.08 [37] and excluding the non-GPU overhead for servers, we estimate the total training energy consumption as approximately 30 GWh. Our estimation method follows Meta’s guideline [36] and is conservative, as it excludes the substantial non-GPU energy overheads (e.g., CPUs) associated with server operations.

## A.5 Average Emission for Each LA-NYC Round Trip by Car

We use the 2023 national average emission rate for light-duty vehicles (gasoline) provided by the U.S. Department of Transportation [104]. The emission rate accounts for tailpipe exhaust, tire wear and brake wear. Specifically, the average PM<sub>2.5</sub> emission rate is 0.008 grams/mile (including 0.004 grams/mile for exhaust, 0.003 grams/mile for brake wear, and 0.001 grams/mile for tire wear), and the average NO<sub>x</sub> emission rate is 0.199 grams/mile for exhaust. We see that half of PM<sub>2.5</sub> for light-duty vehicles comes from brake and tire wear (0.004 gram/miles), which are also produced by other types of vehicles including electric vehicles. The distance for a round-trip between Los Angeles, California, and New York City, New York, is about 5,580 miles. Thus, the average auto emissions for each LA-NYC round trip are estimated as 44.64 grams of PM<sub>2.5</sub> and 1110.42 grams of NO<sub>x</sub>.

## A.6 State-wide Electricity Consumption by U.S. Data Centers in 2023

We show in Fig. 10 the state-wide data center electricity consumption in 2023 [5]. It can be seen that Virginia, Texas and California have the highest data center electricity consumption in 2023. The total national electricity consumption reported by EPRI is slightly lower than the values in [4], and we scale it up accordingly in our calculations to ensure consistency.

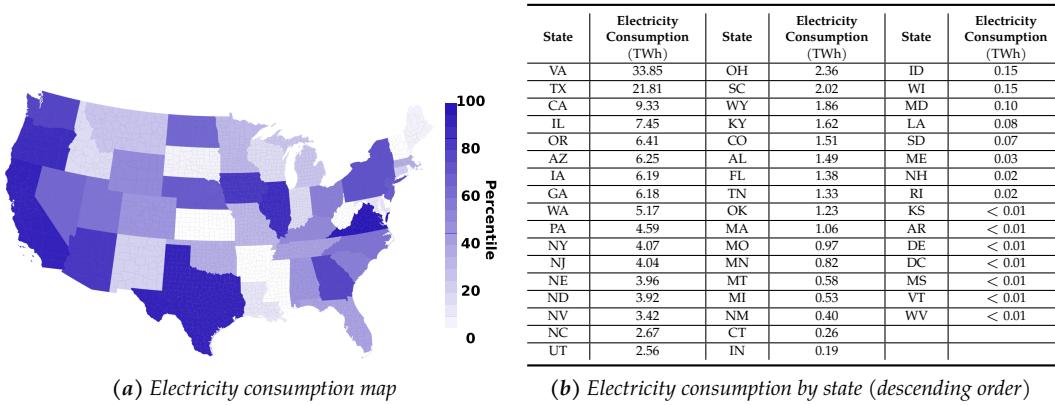


Figure 10: State-level electricity consumption of U.S. data centers in 2023 [5].

## B Additional Results for Health-Informed GLB

### B.1 Details of the experiment setup

We use Meta’s electricity consumption for each U.S. data center location in 2023 [37] for our experiments. Table 7 summarizes the baseline annual energy load  $W_i$  for each data center  $i$ . Since data centers are mostly stable loads in practice, the hourly workload for each location is calculated as  $l_i = \frac{W_i}{T}$ , where  $T$  represents the total number of hours in the study period. The total hourly workload is then computed as  $M_t = \sum_{i=1}^N l_i$ .

We use the annual average industrial electricity prices in different states provided by the EIA [109]. The health price  $p_{i,t}^h$  (\$/MWh) and carbon emission rate  $r_{i,t}^c$  (ton/MWh) are based on data provided by WattTime [77]. WattTime divides the U.S. into more than 100 regions and provides marginal health prices and carbon emission rates for each region. These values are updated every 5 minutes.

**Table 7:** Information about Meta's U.S. data center locations in 2023

Location	Energy (MWh)	Electricity Price $p^e$ (\$/MWh)	Health Price $p^h$ (\$/MWh)	Carbon Intensity $r^c$ (ton/MWh)
Huntsville, AL	614198	71.0	49.49	0.59
Stanton Springs, GA	968565	68.8	19.35	0.58
DeKalb, IL	138965	82.0	51.72	0.58
Altoona, IA	1243306	69.1	57.86	0.65
Sarpy, NE	1148091	76.3	36.72	0.56
Los Lunas, NM	1110100	57.5	18.96	0.68
Forest City, NC	507068	71.5	54.64	0.66
New Albany, OH	793063	70.3	54.32	0.61
Prineville, OR	1375321	75.2	13.31	0.65
Gallatin, TN	116520	62.3	49.49	0.59
Fort Worth, TX	1029570	66.0	44.56	0.51
Eagle Mountain, UT	787740	69.9	21.93	0.74
Henrico, VA	805061	89.2	54.37	0.61

In our study, we use the WattTime data from 0:00 on January 1, 2023, to 23:55 on December 31, 2023. The actual time interval for our experiment is 1 hour, where the 5-minute data points are averaged hourly to compute  $p_{i,t}^h$  and  $r_{i,t}^c$ ,  $\forall t = \{1, 2, \dots, T\}$ , with  $T = 8760$ .

## B.2 Sensitivity analysis of capacity slackness $\lambda$

The parameter  $\lambda \geq 1$  represents the capacity slackness to accept additional loads: The greater  $\lambda$ , the more spatial flexibility there is. Here, we vary  $\lambda$  and report the results for  $\lambda = 1.2$  and  $\lambda = 2.0$  in Tables 8 and 9, respectively. Similar to Table 4, the potential reduction in health costs is significant. An increase in  $\lambda$  corresponds to an expansion in the spatial flexibility, leading to a larger reduction in health cost. Specifically, when  $\lambda = 2.0$ , the health-informed algorithm achieves a reduction in health cost of over 40% compared to the baseline. In contrast, under  $\lambda = 2.0$ , the pure carbon-aware algorithm, which focuses exclusively on optimizing carbon emissions with an infinite carbon price  $p^c = \infty$ , can achieve a reduction in carbon emissions of approximately 12% relative to the baseline. This comparison highlights the great potential for health cost reduction and reinforces the importance of designing HI-GLB.

**Table 8:** Comparison between health-informed and carbon-aware GLB ( $\lambda = 1.2$ )

Metric	Baseline	Carbon-Aware GLB				Health-Informed GLB	
		$p^c = \$0/\text{ton}$	$p^c = \$5/\text{ton}$	$p^c = \$200/\text{ton}$	$p^c = \$\infty/\text{ton}$	HI-GLB ( $p_{i,t}^c = 0$ )	HI-GLB
Health (Million \$)	393.23	374.20 (-4.84%)	374.44 (-4.78%)	395.28 (0.52%)	404.72 (2.92%)	345.26 (-12.20%)	349.29 (-11.17%)
Energy (Million \$)	756.50	732.06 (-3.23%)	732.30 (-3.20%)	747.88 (-1.14%)	761.20 (0.62%)	752.26 (-0.56%)	738.49 (-2.38%)
Carbon (Million Ton)	6.60	6.68 (1.22%)	6.56 (-0.60%)	6.40 (-3.04%)	6.37 (-3.54%)	6.54 (-0.87%)	6.57 (-0.44%)

**Table 9:** Comparison between health-informed and carbon-aware GLB ( $\lambda = 2.0$ )

Metric	Baseline	Carbon-Aware GLB				Health-Informed GLB	
		$p^c = \$0/\text{ton}$	$p^c = \$5/\text{ton}$	$p^c = \$200/\text{ton}$	$p^c = \$\infty/\text{ton}$	HI-GLB ( $p_{i,t}^c = 0$ )	HI-GLB
Health Cost (Million \$)	393.23	368.10 (-6.39%)	409.67 (4.18%)	374.82 (-4.68%)	408.92 (3.99%)	221.38 (-43.70%)	223.96 (-43.05%)
Energy Cost (Million \$)	756.50	702.16 (-7.18%)	702.69 (-7.11%)	723.42 (-4.37%)	767.59 (1.47%)	734.29 (-2.94%)	728.19 (-3.74%)
Carbon (Million Ton)	6.60	6.72 (1.73%)	6.52 (-1.17%)	5.95 (-9.91%)	5.84 (-11.49%)	6.61 (0.19%)	6.55 (-0.71%)