Water Consumption of Data Centers

(ECE 5755) Modern Computer System and Architecture, Final Paper, Fall 2023

Siyu Meng
Cornell Tech
sm2659@cornell.edu

Ruixi (Victoria) Zhang
Cornell Tech
rz442@cornell.edu

Zheng Zhou
Cornell Tech
zz273@cornell.edu

***Abstract*—This paper provides a comprehensive analysis of water consumption in data centers (DCs). We begin by outlining the status quo of water usage in DCs, emphasizing the critical role of water in the cooling system and its consumption in DCs. Then the paper delves into the measurement of water consumption, discussing the established metrics such as Power Usage Effectiveness, Water Usage Effectiveness, and their variations. Furthermore, we explored innovative methods to optimize water consumption in DCs including the selection of cooling systems, the use of renewable energy, and the reduction of water footprint in AI operations. Our analysis aims to provide insights and practices for reducing water consumption in DCs, thereby promoting environmental sustainability in this rapidly growing industry.**

Keywords—Data centers, water consumption, water usage effectiveness, environmentally sustainable computing, water cooling, liquid cooling, power usage effectiveness

I. Background and Motivation

As data centers (DCs) proliferate to accommodate the rising demand of the Internet and Associated Information and Communications Technologies (ICTs), ICT's advancements, and rapid growth of AIs, their environmental impacts grow too (Ristic, 2015). The energy efficiency of DCs has been researched extensively (Ristic, 2015) and the growing carbon footprint of AI models, especially large ones such as GPT-3, has been undergoing public scrutiny (Li, 2023). Unfortunately, however, water consumption of DCs remains to be seriously disregarded. The global AI demand may be accountable for 4.2 – 6.6 billion cubic meters of water withdrawal in 2027, which is more than the total annual water withdrawal of 4 – 6 Denmark or half of the United Kingdom (Li, 2023). This is very concerning, as freshwater scarcity has become one of the most pressing challenges shared by all of us in the wake of the rapidly growing population, depleting water resources, and aging water infrastructures (Li, 2023). DCs require substantial amounts of water, mainly for cooling systems, which poses environmental concerns, especially in water-scarce regions. Understanding and improving water efficiency in DCs is essential for their sustainable operation. This paper aims to address the gap in research on water consumption in DCs. The content focuses on why water is needed and how water is utilized in DCs, assessment of water consumption metrics, and discussion of optimization techniques to promote more sustainable practices for water in DCs.

II. Why do Data Centers Need Water and How Much Water is Consumed

Computing hardware and powerful processors housed by data centers generate immense heat, especially when operating at high clock speeds. High temperatures can lead to thermal throttling, where the system reduces its performance to avoid overheating. To mitigate this, many data centers have transitioned from traditional air cooling, which has low heat dissipation efficiency and will be incapable of handling the chips' thermal design power (TDP) due to increasing transistor counts and the end of Dennard scaling in the near future (Majid, 2021), to more effective liquid cooling solutions. DCs require water/liquid cooling solutions such as cold plates and immersion cooling, to improve heat dissipation capabilities. Thereby they can keep the servers cooler and operate server components beyond the normal frequency range (i.e., overclocking them) (Majid, 2021).

DCs' thirst for water leads to significant water consumption. In 2021, Google's global data centers consumed 4.3 billion gallons of water - on average 450,000 gallons of water on a daily basis. (Heather, 2022) Other companies show similar consumption levels. Moreover, with the advent of advanced AI systems, like GPT, and the corresponding massive GPU training and inference, the demand for water cooling solutions has skyrocketed. Training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly evaporate 700,000 liters of clean freshwater. (Li, 2023) The urgency can also be reflected in part by the recent commitment to "Water Positive by 2030" by increasingly many companies, including Google, Microsoft, and Meta. (Li, 2023) This involves adopting more water-efficient practices, using sustainable water resources, and contributing to community water reuse.

III. How do Data Centers use Water

A. Cooling

Water is used in data centers primarily for on-site cooling of computing servers as mentioned earlier. Cooling methods in DCs include air cooling, water cooling, liquid cooling, etc.

3.1.1 Water cooling (a form of liquid cooling)

3.1.1.1 Cooling towers

Cooling towers generally consist of a closed loop that circulates water within the DCs to maintain server temperatures and an open loop that transfers heat from the chiller to the cooling tower. The closed loop does not consume water, but the open loop involves water evaporation in the cooling tower to disperse heat. This water, often potable, needs to be replenished due to evaporation and to prevent pipe clogging and bacterial growth. Water withdrawal for cooling towers includes both evaporated and discharged water, with typically 80% being evaporated, categorized as consumption. Depending on factors like water quality and operational settings, data centers can evaporate approximately 1-9 liters of water.

3.1.1.2 Outside air cooling with water assistance

It is used by DCs to cool servers using "free" outside air when the climate is suitable. It involves blowing a large amount of outside air through the servers and then exhausting it outside. Although it is generally more water-efficient than cooling towers, it can lead to significant water usage during hot periods due to water evaporation for additional cooling and humidity control. Some data centers opt for combining outside air cooling with traditional cooling towers.

3.1.2 Liquid Cooling

Liquid cooling can be classified into cold-plate cooling, spray cooling, and immersion cooling. On-chip liquid cooling may be employed for AI servers with high power densities and the heat is transferred directly from servers to the cooling infrastructure. Cold plates have typically been placed on the most power-hungry components. Fluid flows through the plates and piping to remove the heat produced by those components (Majid, 2021). While efficient, each cold plate needs to be specifically designed for each new component. Compared with traditional air cooling methods, liquid cooling has higher heat dissipation efficiency. Liquids conduct heat better than air, remove the heat faster, and thereby reduce the temperature of the device more effectively and improve the life and performance of the device. Liquid cooling has lower noise, more flexible design, and is more environmentally friendly as well.

3.1.2.1 Immersion cooling

With immersion cooling, entire servers are submerged in a tank and the heat is dissipated by direct contact with a dielectric liquid (Majid, 2021). Heat removal can happen In 1PIC (1-phase immersion cooling), the tank liquid absorbs the heat and circulates using pumps, whereas in 2PIC (2-phase immersion cooling) a phase-change process from liquid to vapor (via boiling) carries the heat away. Two-phase immersion cooling (2PIC) is the most promising technology - it can dissipate large amounts of heat (Majid, 2021). Operating server parts at higher frequencies, i.e. overclocking, increases power consumption substantially and immersion cooling provides power savings that partially offset the higher power requirements. Immersion cooling offers high thermal dissipation and low junction temperatures, allowing overclocking for longer periods of time (Majid 2021).

The results of an experiment that evaluates the impact of different cooling methods and overclocking conditions on the lifetime of a 5nm Xeon processor are shown in Table 1. Among all the cooling systems, HFE-7000 immersion cooling (a type of engineered fluid used in cooling systems) achieves the lowest junction temperature, at 51°C, with the longest predicted lifetime of over 10 years. When overclocked, the predicted lifetime of 5 years matches the baseline of air cooling. This effectively demonstrates that immersion cooling can compensate for the lifetime degradation due to overclocking (Majid 2021).

Table 1

PROJECTED LIFETIME COMPARISON FOR RUNNING A XEON PROCESSOR IN AIR AND 2PIC AT NOMINAL AND OVERCLOCKING CONDITIONS.

Cooling	OC	Voltage	Tj Max	DTj	Lifetime
Air cooling	✕	0.90V	85°C	20°-85°C	5 years
Air cooling	✔	0.98V	101°C	20°-101°C	< 1 year
FC-3284	✕	0.90V	66°C	50°-65°C	> 10 years
FC-3284	✔	0.98V	74°C	50°C-74°C	4 years
HFE-7000	✕	0.90V	51°C	35°C-51°C	>10 years
HFE-7000	✔	0.98V	60°C	35°-60°C	5 years

B. Others

With regards to electricity generation, the U.S. national average for water withdrawal at 43.8 liters per kWh and consumption at 3.1 liters per kWh. Meta, for example, reported an average water consumption of 3.58 liters per kWh in 2022. In AI chip and server manufacturing, semiconductor plants require significant amounts of clean water for cooling and production processes. These facilities often have a low water recycling rate, with an average of 23% in Singapore, highlighting the need for improvement in water conservation practices.

IV. How to Measure Water Consumption

A. PUE

Before delving into the discussion of water consumption measurement, it is pertinent to introduce a prominent metric utilized in the contemporary industry for estimating the energy efficiency of infrastructure within data centers (Avelar, Azevedo, French, & Power, 2012). This metric is commonly referred to as Power Usage Effectiveness (PUE), which is calculated as follows:

$$\boldsymbol{PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy}}$$

The equation has the following components:

- **IT Equipment Energy**: This includes the energy consumed by equipment dedicated to the management, storage, or routing data within the computing space.
- **Total Facility Energy**: This incorporates not only energy consumed by IT equipment energy but also

includes the energy utilized to power the IT equipment and maintain the overall functionality of the data center such as cooling system components and data center lighting.

The PUE metric calculates the ratio of total energy consumption to IT equipment energy consumption, the latter being fundamental to the primary functionality of the data center. Ideally, a PUE value of one is sought, signifying that all energy consumed by the data center is directed towards powering IT equipment to support and optimize the computing system. In the real world, most data centers have PUE value around 3.0 or greater. However, with the proper design, the performance of PUE could achieve 1.6 or better (Avelar, Azevedo, French, & Power, 2012).

B. WUE

Water Usage Effectiveness (WUE) serves as an extension of PUE, focusing on evaluating the water sustainability aspects within a data center. WUE incorporates two distinct sub-metrics: site-based and source-based. These sub-metrics provide a framework for assessing the water utilization efficiency of a data center, addressing both on-site water management and the origin of the water resources involved in the facility's operations.

a) Formula of Site-based WUE:

$$\text{WUE} = \frac{\text{Annual Water Usage}}{\text{IT Equipment Energy}}$$

The equation has the following components:

- **Annual Water Usage:** It only assesses the water used on-site for operation of the data center, which includes water used for humidification, water evaporated on-site, cooling of the data center and its support system.
- **IT Equipment Energy:** Same denominator in PUE formula in section 3.1

b) Formula of Source-based WUE:

$$\text{WUE}_{\text{source}} = \frac{\text{Annual Source Energy Water Usage} + \text{Annual Site Water Usage}}{\text{IT Equipment Energy}}$$

The equation has the following components:

- **Annual Source Energy Water Usage:** It includes water used off-site in the production of energy, such as generating electricity for power IT equipment.
- **Annual Site Water Usage:** Same as annual water usage in site based WUE.
- **IT Equipment Energy:** Same denominator in PUE formula in section 3.1

The units of WUE are liters/kilowatt-hour (L/kWh) (Christian, 2010). This metric serves as a quantifiable measure to assess the volume of water consumed by a data center in relation to the power required for IT Equipment operations. The objective is to minimize the WUE value, ideally reaching 0.0, signifying an absence of water usage in the operational processes of the data center. However, practical considerations, specifically attributed to the imperative role of cooling system components and the prevalence of water-intensive electricity generation, prevent the realization of a WUE value of zero.

Due to different components of site-based and source-based WUE, they have distinct and relevant applications. Source based WUE is intricately linked to site planning and data center design. Decisions such as selecting a cool and dry site, as opposed to a hot and humid one, contribute to the reduction of off-site water usage, particularly at power-generation stations. Conversely, Site-based WUE exclusively estimates on-site water usage, a metric easily measurable. Consequently, this data predominantly serves the purpose of benchmarking, enabling comparisons among major corporations in their endeavors to optimize data center operations in response to environmental dynamics. The data derived from site based WUE offers insights into a data center's efficiency concerning its cooling systems, humidification, and water evaporation. This statistical information empowers companies to scrutinize individual components, thereby strategically reducing the overall WUE value. However, site based WUE possesses limitations, primarily in its exclusive consideration of on-site water consumption. It overlooks off-site compensatory measures implemented to offset reduced on-site consumption. For instance, altering or enhancing the design of the cooling system to diminish on-site water consumption might necessitate increased electricity usage, with the majority of electricity generation processes remaining water intensive. The site based WUE, therefore, fails to address this intricate interplay and may lead to misleading conclusions.

In the report for data center water usage from the year 2021, two prominent companies, AWS and Meta, exhibited noteworthy figures in contrast to the industry average of 1.8. Specifically, AWS recorded a WUE of 0.25, while Meta reported a figure of 0.26 (Zhang, 2022). These values indicate a significant deviation from the industry norm, suggesting a marked efficiency in water usage management by both companies. To diminish on-site water usage, AWS implemented two distinct cooling systems – a direct evaporating system and a free-air cooling system. In the former, hot air is drawn from the external environment and circulated through water-soaked cooling pads. The ensuing evaporation process effectively reduces the air's temperature, and the cool air is subsequently directed into the server rooms. The latter system involves the installation of sensors within AWS data centers that monitor meteorological parameters such as temperature and humidity. When conditions fall within a predefined safe operating range, the evaporative cooling system deactivates, allowing cool external air to be drawn into the server rooms. Both cooling systems, designed to substitute water usage with air-based mechanisms, markedly contribute to the reduction of on-site water consumption.

	WUE (L/kWh)
Industry Avg	1.80
AWS	0.25
Meta	0.26

Figure 1. 2021 WUE data

Figure 2. 2017 Meta WUE trend

Beyond the imperative of reducing on-site water usage, both Meta and AWS have implemented multifaceted strategies to enhance their water positive percentages. The concept of water positivity emphasizes the efficient utilization of water in data center operations and a commitment to returning more water to the environment than is consumed. This is achieved through initiatives supporting local communities and water restoration projects. Meta, for instance, has collaborated with local organizations, embarking on a significant investment in 25 water restoration projects across seven watersheds where their data centers have been operational since 2017 (Loher, 2023). This strategic engagement underscores Meta's commitment to not only reduce its impact on local water resources but actively contribute to the restoration and sustenance of these ecosystems. Similarly, AWS has undertaken initiatives aimed at enhancing its water positive standing. This includes the adoption of sustainable water resources, such as recycled water, in its data center operations. Moreover, AWS has actively participated in returning water for community reuse and similar water restoration projects (Zhang, 2022). These concerted efforts by Meta and AWS in augmenting their water positive percentages exemplify a proactive stance toward environmental sustainability and corporate responsibility. Beyond mere reduction in water consumption, such initiatives contribute substantively to the overall health and resilience of local ecosystems, fostering a balanced and mutually beneficial relationship between data center operations and the communities in which they are situated.

C. WSUE

The Water Scarcity Usage Effectiveness (WSUE) metric is an important tool for assessing the impact of data center operations on regional water availability, taking into account both direct and indirect water usage and location based water scarcity level (Chen & Wemhoff, 2022). A higher WSUE value signifies a greater environmental impact in areas of water scarcity. The metrics is calculate as follows:
The equation has the following components:

- **WSF (Water Scarcity Footprint):** This quantifies the potential environmental impact related to water, calculated as Water Consumption (in cubic meters) multiplied by the Water Scarcity Indicator (WSI), which evaluates the level of water scarcity in a specific region.
- **Pit (Primary Infrastructure Technology):** This focuses on the core technological infrastructure of the data center and its influence on water usage.
- **AWARE CF (Available Water Remaining, Characterization Factor):** This measures the potential environmental impact of water consumption based on local water scarcity. It is calculated as AMD_ref (the reference value of water availability minus demand) divided by AMD (actual water availability minus demand). High AWARE CF values indicate water-scarce regions, as illustrated in Figure 3, which shows high values in counties in New Mexico, Texas, and Arizona (McMullen & Wemhoff, 2023). However, the WSUE metric faces limitations due to the range of AWARE CF being capped between 0.1 and 100, which might cause artificially lower WSUE values in water-scarce areas. This limitation could potentially understate the significance of on-site water usage within the metric (Chen & Wemhoff, 2022).
- **WUE (Water Usage Effectiveness):** A standard metric for assessing water efficiency in data centers, discussed in section 3.2.
- **SWI (Scarce Water Index):** Measured in liters per kilowatt-hour (L/kWh), quantifies the impact of electricity consumption on water availability, with high values indicating that indirect water use is significantly affecting water resources.
- **PUE (Power Usage Effectiveness):** A common sustainability metric for data centers, discussed in section 3.3

The WSUE formula integrates water usage efficiency (WUE) and energy efficiency (PUE) with local water scarcity factors (AWARE CF and SWI) to holistically assess the environmental impact of data centers. It considers both direct water usage and indirect water usage through electricity consumption, making it a more comprehensive metric. The formula adapts to specific locations by using county-specific AWARE CF and region-specific SWI, reflecting the local

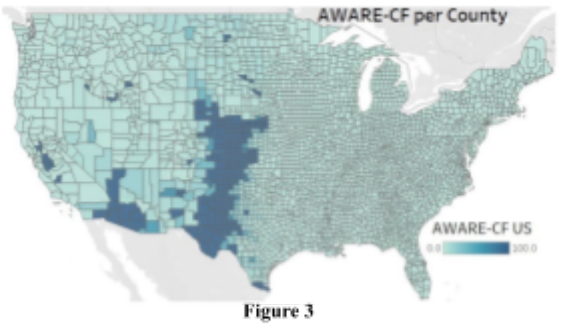

Figure 3

water availability and electricity flows. This approach helps in selecting data center sites more sustainably, especially in water-stressed areas, by emphasizing both water and energy efficiency in a region-specific context.

D. Blowdown

The environmental impact of data centers on water encompasses not just water consumption but also wastewater production, particularly on the process known as blowdown. This aspect is critical in the operation of cooling towers within data centers, involving the selective removal and replacement of cooling water (Sharma, 2010). This process is essential for preventing the excessive buildup of undesirable components, maintaining the efficiency and longevity of the cooling systems. Blowdown of datacenter can be calculated using the equation below (Siddik, Shehabi, & Marston, 2021):

$$R_{\text{Blowdown}} = \frac{1}{C-1} \times R_{\text{Evaporation}}$$

The equation has the following components:

- **R_Blowdown:** blowdown rate required for a cooling tower(in cubic meters per megawatt-hour, $m^3\ MWh^{-1}$).
- **C:** cycle of concentration for dissolved solids
- **R_Evaporation:** rate of evaporation, which is the amount of water that evaporates from the cooling process.

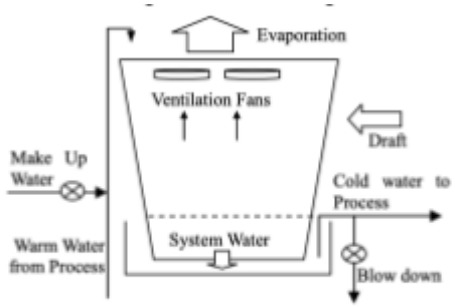

Figure 4. Blowdown of DC cooling process

The main factor that influences the blowdown rate in data center cooling towers is the quality of the makeup water – the fresh water added to the system, with varying levels of minerals and impurities in different water sources impacting the process (Beaty, Quirk, & Morrison, 2019). To optimize the blowdown process and reduce its environmental impact, data centers can adopt several strategies. Improving the water treatment program is a key approach, as a more effective treatment can better manage the concentration of solids and reduce the need for frequent blowdowns. Additionally, employing more corrosion-resistant materials in the construction of cooling towers can diminish the impact of water chemistry on the system, thereby optimizing the blowdown rate (Beaty, Quirk, & Morrison, 2019).

V. Water Consumption Optimization Strategies

A. Selection of Cooling System

In a computational modeling study using the Villanova Thermodynamic Analysis of Systems (VTAS), a flow network modeling tool, the thermal dynamics, fluid mechanics, and heat transfer characteristics of different data center cooling systems were analyzed (Chen & Wemhoff, 2022). As illustrated in Figure 5, the study compared three cooling

Figure 5. three types of cooling mechanism

systems' water consumption.

- CRAC operates similarly to a mechanical refrigeration system. It pumps cold air through perforated tiles to cool the computers. The airflow direction causes the cold air to push the hot air towards the hot aisle opposite the computers. It has high energy consumption due to the use of a compressor, but requires less water.
- CRAH systems include chillers and cooling towers. It takes hot air in a room and sends it to a cooling coil filled with chilled water. This system requires high water consumption and produces high wastewater (blowdown) due to the cooling tower. However, it is more energy-efficient.
- The pure evaporative cooling system expels all air from the data center, relying on natural evaporation without a compressor, resulting in high cooling efficiency with low water and energy consumption. However, this method can only be used in dry climates, limiting its geographical and seasonal viability.

The study considered diverse water scarcity levels across locations like Boston, Denver, Miami, Phoenix, and San Francisco. The result is demonstrated in Table 2. The study's findings highlight the suitability of different cooling methods for data centers based on regional characteristics, particularly focusing on water scarcity and climate.

- Evaporative cooling solution is most effective, with lowest WSUE PUE, and WUE compared to other cooling methods, but it is only in Denver and San Francisco. This method is advantageous due to its airside economization and the absence of a compressor, making it ideal for dry climates with low humidity. Specifically In San Francisco, the evaporative cooling solution has a higher WSUE (124.86 L/kWh) than in Denver (33.40 L/kWh), due to San Francisco's higher SWI (123.6 L/kWh), highlighting the impact of indirect water consumption on water stress.

Table 2. Performance Metrics of DC with Different Cooling System in Various Regions.

Location	AWARE CF (0-100)	SWI (L/kWh)	Cooling System	PUE	WUE (L/kWh)	WSUE (L/kWh)
Boston	0.27	0.53	CRAC cooling	1.69	0	0.89
			CRAH cooling	1.52	2.11	1.38
			Evaporative cooling	1.01	0.01	0.54
Denver	100	24.16	CRAC cooling	1.69	0	40.83
			CRAH cooling	1.52	2.14	250.72
			Evaporative cooling	1.01	0.09	33.40
Miami	0.67	0.91	CRAC cooling	1.86	0	1.69
			CRAH cooling	1.61	2.37	3.04
			Evaporative cooling	1.01	0.01	0.92
Phoenix	100	890.3	CRAC cooling	1.88	0	1,673.76
			CRAH cooling	1.6	2.44	1,668.48
			Evaporative cooling	1.01	0.30	929.20
San Francisco	0.96	123.6	CRAC cooling	1.71	0	211.36
			CRAH cooling	1.53	2.17	191.19
			Evaporative cooling	1.01	0.02	124.86

*The modeled evaporative cooling system supply air temperatures in Boston, Miami, and Phoenix are too high for practical comparison of metrics with other cooling systems.

- CRAH-based cooling systems, though more water-intensive, are preferable in areas like Phoenix and San Francisco, where they can achieve lower WSUE values compared to CRAC systems, especially in regions with high PUE and/or SWI values. These systems are more energy-efficient due to their use of chillers and cooling towers, which reduce indirect water use.
- For cities like Denver, which experience high water scarcity, CRAC-based cooling is more suitable. It shows a lower WSUE in such environments, making it a better choice over CRAH systems where direct water use has a more significant impact than indirect water use.

Thus, the choice of cooling systems for data centers should be informed by a thorough understanding of local environmental conditions, specifically water availability and climate, to ensure optimal efficiency and minimal environmental impact.

B. Use of Renewable Energy

Another group of researchers at Villanova University examined the impact of incorporating 25% renewable energy, specifically wind and solar, into data centers on Water Scarcity Usage Effectiveness (WSUE). Figures 6 and 7 in their publication visually demonstrated the experimental results. A key discovery was the regional differences in WSUE following renewable energy integration. Counties with improved WSUE, marked in green, showed benefits from renewable energy, while those with worsened WSUE, indicated in red, highlighted the varied regional effects of renewable energy on water scarcity(McMullen & Wemhoff, 2023).

The study observed that the greatest benefits of renewable energy in data centers were in areas with high ratios of Scarce Water Index (SWI) to Available Water Remaining Characterization Factor (Acf). This ratio indicates the comparison between scarce water usage due to grid power and on-site generation. Figure 8 illustrates the geographic distribution of the SWI/Acf ratio and its correlation with WSUE improvement, showing a similar pattern. Additionally, Figure 9 highlights a clear positive correlation between the percent improvement in WSUE and the SWI/Acf ratio, emphasizing the link between water scarcity and renewable energy benefits in varying regions.

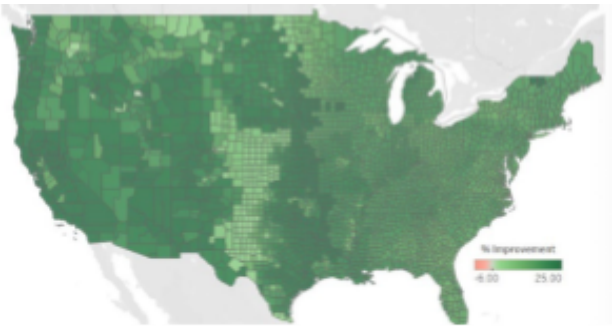

Figure 7: Percent improvement in WSUE after implement 25% on-site solar energy production

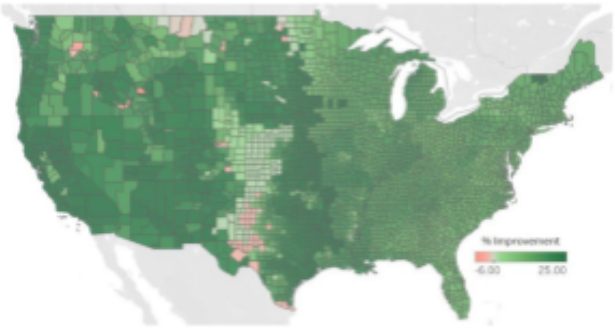

Figure 6: Percent improvement in WSUE after implement 25% on-site solar energy production

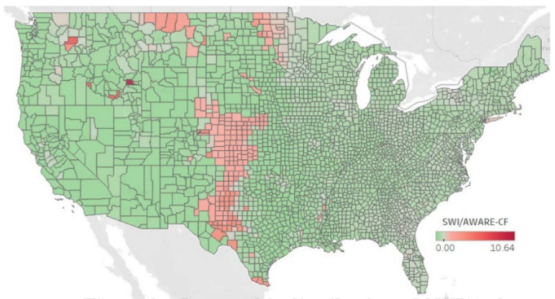

Figure 8: Geographic distribution of SWI/Acf

In counties with high SWI/Acf ratios, renewable energy significantly benefits data centers. This is because renewable sources like wind or solar have lower water footprints compared to traditional energy, impacting indirect water use. Direct water use, affecting the Acf, is the balance of consumption against regional availability. Indirect use, influencing the SWI, reflects how traditional energy depletes water resources. In areas with high SWI, switching to renewables reduces the SWI by lowering indirect water use. However, in areas with a low SWI/Acf ratio (high AWI), where direct water use is predominant, renewable energy has less impact on WSUE, as the issue is more related to direct water consumption than to electricity-induced water use.

Another key finding was the greater improvement in WSUE with on-site wind power compared to solar power. This improvement was attributed to the lower Energy Water Intensity Factor (EWIF) associated with wind power. The EWIF quantifies the water used to produce energy, underscoring the efficiency and environmental advantages of wind power in data center operations.

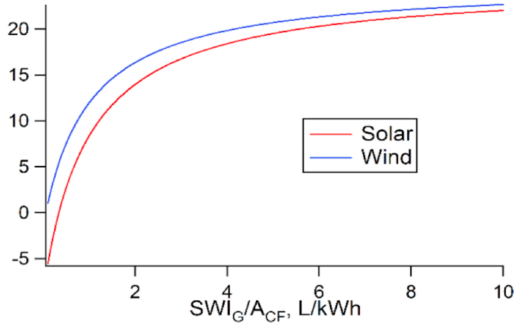

Figure 9: SWI/Acf to WSUE Improvement

C. *Flexible nature of AI workloads can help minimize AI's water footprint*

Deciding "when" and "where" to train a large AI model can significantly affect the water footprint (Li, 2023). WUE of data centers varies both spatially and temporally, indicating that water efficiency is subject to geographical and time-based diversity. In Figure 10, the dashed line represents a linear regression model, showing that the eGRID-level scope 2 carbon emission (indirect emissions associated with the purchase of electricity, steam, heat, or cooling) and water consumption efficiencies are not aligned. Figure 11 is a 5-day snapshot of scope-2 carbon emission rate and water consumption intensity for electricity generation serving Virginia, starting from April 4, 2022 (Li, 2023). These two figures show that the water efficiency has spatial-temporal diversity - on-site water efficiency changes due to variations of outside weather conditions (Li, 2023).

By dynamically scheduling AI model training and inference, it is possible to reduce the water footprint. AI training and inference possess spatial flexibility, meaning they can be processed in various data centers without significantly affecting latency due to recent advances in data center networking. This allows for strategic placement in regions with better water efficiency. There is also temporal flexibility, as AI models don't need continuous training but can be scheduled intermittently.

For example, the training task of a small AI model can be scheduled at midnight in a data center location with better water efficiency. Water-conscious users can utilize the inference services of AI models during hours and in data centers that are more water-efficient, thereby reducing AI's water footprint (Li, 2023). The strategy of 'unfollowing the sun' involves avoiding the high-temperature hours of a day when WUE is high (Li, 2023). As a result, scheduling tasks in cooler periods or locations can avoid peak temperatures that require more cooling and, consequently, more water.

Performance flexibility is also achievable through techniques like model pruning and compression, which can reduce the resources needed for AI without significantly impacting performance.

More transparency in run-time water efficiency and detailed water usage information for cooling and electricity generation are needed from data center operators to enable more sustainable AI.

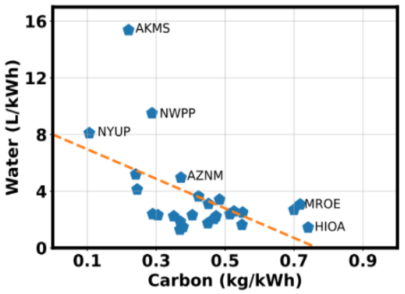

Figure 10: eGRID-level carbon/water efficiency

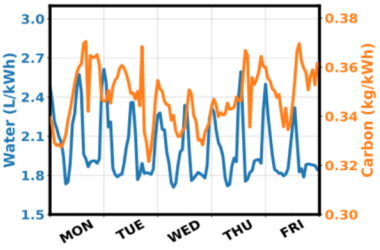

Figure 11: Hourly carbon/water efficiency

VI. Conclusion

Our findings underscore a noteworthy trend within the corporate landscape, indicating a growing commitment among certain companies to integrate water conservation measures alongside technological advancements. These forward-thinking entities are actively engaged in the innovation of new techniques and machinery, demonstrating a conscientious effort to curtail water consumption or explore alternatives, including the integration of renewable energy sources. However, it is imperative to acknowledge a concerning aspect observed in our study—there exist companies that either abstain from disclosing their annual water consumption data or neglect to account for water usage within their data center operations. This opacity raises questions about the overall industry's commitment to sustainable practices, particularly concerning water resource management. This research serves as a clarion call to all industries, urging them to scrutinize their water usage practices and instigate comprehensive evaluations aimed at enhancing their water positive percentage.

References

[1] Avelar, V., Azevedo, D., French, A., & Power, E. N. (2012). PUE: a comprehensive examination of the metric. White paper, 49.

[2] Barroso, L. A., Hölzle, U., & Ranganathan, P. (2019). The datacenter as a computer: Designing warehouse-scale machines. Springer Nature

[3] Beaty, D. L., P.E., Quirk, D., P.E., & Morrison, F. T. (2019). Designing data center waterside economizers. ASHRAE Journal, 61(1), 57-61. Retrieved from https://www.proquest.com/scholarly-journals/designing-data-center-waterside-economizers/docview/2313379292/se-2

[4] Chen, Li & Wemhoff, Aaron. (2022). Characterizing Data Center Cooling System Water Stress in the United States.

[5] Christian, B. E. L. A. D. Y. (2010). A Green Grid Data Center Sustainability Metric. The Green Grid white paper, 32.

[6] Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less" Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. arXiv preprint arXiv:2304.03271.

[7] McMullen, M, & Wemhoff, AP. "Data Center Environmental Burden Reduction Through On-Site Renewable Power Generation." Proceedings of the ASME 2023 17th International Conference on Energy Sustainability collocated with the ASME 2023 Heat Transfer Summer Conference. ASME 2023 17th International Conference on Energy Sustainability. Washington, DC, USA. July 10–12, 2023. V001T02A015. ASME. https://doi.org/10.1115/ES2023-107496

[8] Mytton, D. Data centre water consumption. npj Clean Water 4, 11 (2021). https://doi.org/10.1038/s41545-021-00101-w

[9] Sharma, R, Shih, R, McReynolds, A, Bash, C, Patel, C, & Christian, T (2010). "Water Utilization in Data Center Infrastructure." Proceedings of the ASME 2010 International Mechanical Engineering Congress and Exposition. Volume 5: Energy Systems Analysis, Thermodynamics and Sustainability; NanoEngineering for Energy; Engineering to Address Climate Change, Parts A and B. Vancouver, British Columbia, Canada. November 12–18, 2010. pp. 1413-1419. ASME. https://doi.org/10.1115/IMECE2010-40819

[10] Siddik, M. A. B., Shehabi, A., & Marston, L. (2021). The environmental footprint of data centers in the United States. Environmental Research Letters, 16(6), 064017.

[11] Zhang, M. (2023, August 4). Data Center water usage: Billions of gallons every year. Dgtl Infra. https://dgtlinfra.com/data-center-water-usage/

[12] Loher, N. (2023, March 20). What does it mean to be “water positive”? Meta Sustainability. https://sustainability.fb.com/blog/2023/03/15/what-does-it-mean-to-be-water-positive/#:~:text=Striving%20for%20Water%20Positive%20and,address%20local%20needs%20and%20context.

[13] Ristic B, Madani K, Makuch Z. (2015) The Water Footprint of Data Centers. Sustainability; 7(8):11260-11284. https://doi.org/10.3390/su70811260

[14] Majid J, Ioannis M, Íñigo G, Pulkit M, Ashish R, Husam A, Bharath R, Phillip T, Christian B, Marcus F, Ricardo B. (2021) "Cost-Efficient Overclocking in Immersion-Cooled Datacenters," in 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 2021 pp. 623-636. https://doi.ieeecomputersociety.org/10.1109/ISCA52012.2021.00055

[15] Heather, C. (2022) Sip or guzzle? Here's how Google's data centers use water, Practical Magic, Greenbiz, November 2022. https://www.greenbiz.com/article/sip-or-guzzle-heres-how-googles-data-centers-use-water#:~:text=Here's%20the%20overall%20number%20to,water%20on%20a%20daily%20basis.

[16] Meta. Sustainability report, 2023, https://sustainability.fb.com/ 2023-sustainability-report/.