Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

Pengfei Li — *UC Riverside*
Jianyi Yang — *UC Riverside*
Mohammad A. Islam — *UT Arlington*
Shaolei Ren[1] — *UC Riverside*

Abstract

The growing carbon footprint of artificial intelligence (AI) models, especially large ones such as GPT-3 and GPT-4, has been undergoing public scrutiny. Unfortunately, however, the equally important and enormous water footprint of AI models has remained under the radar. For example, training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly consume **700,000 liters** of clean freshwater (enough for producing **370 BMW cars** or **320 Tesla electric vehicles**) and the water consumption would have been **tripled** if training were done in Microsoft's Asian data centers, but such information has been kept as a *secret*. This is extremely 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. To respond to the global water challenges, AI models can, and also should, take social responsibility and lead by example by addressing their own water footprint. In this paper, we provide a principled methodology to estimate fine-grained water footprint of AI models, and also discuss the unique spatial-temporal diversities of AI models' runtime water efficiency. Finally, we highlight the necessity of holistically addressing water footprint along with carbon footprint to enable truly sustainable AI.

Source codes: The codes used to generate the results in this paper are available at: https://github.com/Ren-Research/Making-AI-Less-Thirsty

1 Introduction

- "Water is a finite resource, and every drop matters." — *Facebook (now Meta) Sustainability Report 2020* [1].
- "At AWS, we know that water is a precious resource." — *Amazon's Water Stewardship 2023* [2].
- "Fresh, clean water is one of the most precious resources on Earth ... Now we're taking urgent action to support water security and healthy ecosystems." — *Google's Water Commitment 2023* [3].
- "Water is a human right and the common development denominator to shape a better future. But water is in deep trouble." — *U.N. Secretary-General António Guterres at the U.N. Water Conference 2023* [4].
- "Historic droughts threaten our supply of water ... As the source of both life and livelihoods, water security is central to human and national security." — *U.S. White House Action Plan on Global Water Security 2022* [5].
- · · ·

Artificial intelligence (AI) models have witnessed remarkable breakthroughs and success in numerous areas of critical importance to our society over the last decade, including in the ongoing combat against several global challenges such as climate changes [6]. Increasingly many AI models are trained and deployed on power-hungry servers housed inside warehouse-scale data centers, which are often known as energy hogs [7]. Consequently, the environmental footprint of AI models, in particular carbon footprint, has been undergoing public scrutiny, driving the recent progress in AI carbon efficiency [8–11]. Unfortunately, however, the enormous water footprint of AI models — millions of liters of clean freshwater *consumed* for generating electricity to power data center servers and for cooling these servers — has remained under the radar, which, if not properly addressed, can become a major roadblock for socially responsible and environmentally sustainable evolution of future AI.

1.1 Motivation

Despite the water cycle through our planet's natural ecosystem, clean freshwater resource available and suitable for use is extremely limited and unevenly distributed across the globe. In fact, freshwater scarcity

[1]Corresponding author: Shaolei Ren (sren@ece.ucr.edu)

is one of the most pressing challenges shared by all of us in the wake of the rapidly growing population [12,13]. Severe water scarcity has already been affecting 4 billion people, or approximately two-thirds of the global population, for at least one month each year [13,14]. Without integrated and inclusive approaches to addressing the global water challenge, nearly half of the world's population will endure severe water stress by 2030 [5], and roughly one in every four children worldwide will be living in areas subject to extremely high water stress by 2040 [14].

Warehouse-scale data centers — physical "homes" where the majority of AI models, especially large ones like GPT-3 and GPT-4, are physically trained and deployed — are known to be energy-intensive, collectively accounting for 2% of the global electricity usage and large carbon footprint [7, 8]. Nonetheless, what is much less known is that data centers are also extremely "thirsty" and consume an enormous amount of clean freshwater [15,16]. For example, even excluding water consumption in leased third-party colocation facilities, Google's self-owned data centers in the U.S. alone consumed **12.7 billion** liters of freshwater for on-site cooling in 2021, roughly 90% of which was potable water [17]. This astonishing amount of water is enough for producing **6.9 million BMW cars**, or **5.7 million electric vehicles** (including battery cell manufacturing) by Tesla at its Gigafactory in Berlin, according to water consumption data released by BMW [18] and Tesla [19]. The combined water footprint of U.S. data centers altogether in 2014 was estimated at 626 billion liters [7].

Figure 1 shows that 37.03% of the U.S. area can be under severe drought or worse, where thousands of data centers are located [20,21]. The presence of megawatt data centers clearly results in a huge environmental impact on regional water systems. Additionally, due to the aging public water infrastructure, the need for water conservation remains equally important, even in non-drought areas. For example, in Florida and Singapore, the freshwater shortage is still a key challenge due to the water infrastructure constraints [22,23]. Moreover, it is extremely costly to expand the aging public water infrastructure that is already operating near limits in many parts of the world. The addition of water-thirsty data centers to accommodate new AI model development can certainly worsen the situation.

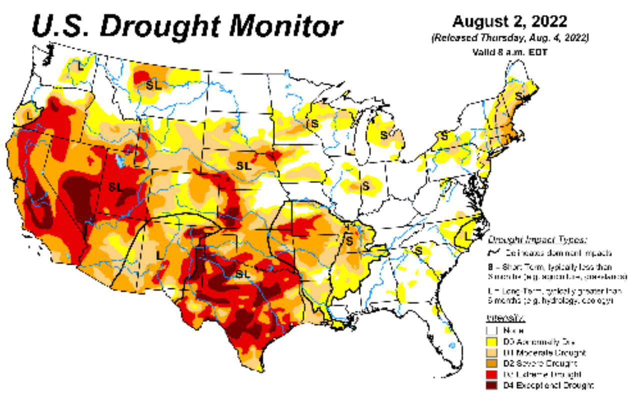


Figure 1: *US drought map for August 2, 2022, with 4.47% area under exceptional drought (D4), 18.96% area under extreme drought or worse (D3-D4), and 37.03% area under severe drought or worse (D2-D4) [20].*

As one of the most prominent and exponentially expanding workloads in data centers [8,24,25], AI models can, and also should, take social responsibility and lead by example in the collective efforts to combat the global water scarcity challenge by cutting their own water footprint. Despite its profound environmental and societal impact, however, the enormous water footprint of AI models has remained hidden from the AI community as well as the general public. Therefore, it is truly a critical time to uncover and address the AI model's secret water footprint amid the increasingly severe freshwater scarcity crisis, worsened extended droughts, and quickly aging public water infrastructure. The urgency can also be reflected in part by the recent commitment to "*Water Positive by 2030*" by increasingly many companies, including Google [17], Microsoft [26], Meta [27] and Amazon [2].

1.2 Overview of Our Study

Recognizing the enormous water footprint as a critical concern for socially responsible and environmentally sustainable AI, we make the first-of-its-kind efforts to uncover the secret water footprint of AI models. Let us first take the GPT-3 model for language services as a concrete example [28].

• **Training:** Training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly consume **700,000 liters** of clean freshwater, enough for producing **370 BMW cars** or **320 Tesla electric vehicles**, and these numbers would have been **tripled** if GPT-3 were trained in Microsoft's Asian data centers, according to public data sources [18, 19, 29, 30].[2] If training a similar large AI model elsewhere with average water efficiency of 3.8L/kWh [31], the on-site water consumption can be as much as 4.9 million liters, enough for producing roughly 2,600 BMW cars or 2,200 Tesla electric vehicles. Moreover, training GPT-3 is also responsible for an additional off-site water footprint of **2.8 million liters** due to electricity usage (assuming water usage efficiency at the U.S. national average level 1.8L/kWh [32] and power usage effectiveness 1.2). Thus, combined together, this would put GPT-3's total water footprint for training at **3.5 million liters** if trained in the U.S., or **4.9 million liters** if trained in Asia.

• **Inference:** ChatGPT needs to "drink" a **500ml bottle of water** for a simple conversation of roughly 20-50 questions and answers, depending on when and where ChatGPT is deployed.[3] While a 500ml bottle of water might not seem too much, the total combined water footprint for inference is still extremely large, considering ChatGPT's billions of users.

All these numbers are likely to increase by multiple times for the newly-launched GPT-4 that has a significantly larger model size. But, up to this point, there has been little public data available to form a reasonable estimate of the water footprint for GPT-4.

Next, by using a principled methodology to estimate the fine-grained water footprint, we show concretely that AI models such as Google's LaMDA [34] can consume a stunning amount of water in the order of millions of liters. We also show that WUE (Water Usage Effectiveness, a measure of water efficiency) is varying both spatially and temporally, implying that judiciously deciding "when" and "where" to train a large AI model can significantly cut the water footprint. In addition, we point out the need of increasing transparency of AI models' water footprint, including disclosing more information about operational data and keeping users informed of the runtime WUE. Finally, we highlight the necessity of holistically addressing water footprint along with carbon footprint to enable truly sustainable AI — *the water footprint of AI models can no longer stay under the radar*.

2 Background

While the embodied water footprint (e.g., manufacturing of servers and GPUs for training AI models) not directly related to AI model training or inference is part of the lifecycle assessment of environmental footprint and can be of independent interest, we focus on the *operational* water footprint of AI models directly associated with training and inference. Thus, in what follows, we introduce the basics of data center infrastructure and then describe the way that data centers (and hence, the hosted AI models) consume water.

Data center power infrastructure. While data center power infrastructures vary from one to another, a common infrastructure design follows the hierarchical tree type. As illustrated in Figure 2, grid/utility power enters the data center through an automatic transfer switch (ATS) which switches over the main power supply to the backup generator in case of grid power failures. The ATS feeds power to a centralized (or distributed in some systems) uninterrupted power supply (UPS), which supplies "protected/conditioned" power to multiple power distribution units (PDUs) that further deliver power to the computer servers. Meanwhile, the cooling system ensures that the servers are not overheated. The computer servers are where AI models are trained and deployed for inference. These servers may also have specialized designs, including multiple GPUs and/or purpose-built hardware, to speed up AI model training and inference [24].

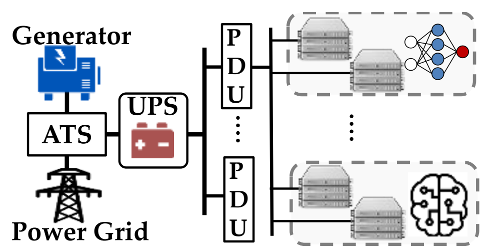


Figure 2: *Data center power infrastructure.*

[2]Without data from the model developer, we set the water usage effectiveness (WUE) as 0.55L/kWh for training in Microsoft's U.S. data centers, and 1.65L/kWh for its Asian data centers based on Microsoft-reported average WUE [30]. The water consumption data for car production is provided by BMW [18] and Telsa [19]. Moreover, in all the estimated water footprints, the overhead for hyperparameter tuning and training failures (if any) is not included.

[3]Without data from the model developer, we assume that each inference consumes $e = 0.00396$kWh energy based on the estimate of [33], the data center's power usage effectiveness (PUE) is 1.2, electricity water intensity factor ($EWIF$) for off-site electricity generation is 1.8L/kWh [32], and water usage effectiveness (WUE_{on}) for on-site cooling is between 0.5L/kWh and 5L/kWh depending on weather conditions. The total water footprint is calculated as $e \cdot [PUE \cdot EWIF + WUE_{on}]$.

Data center water footprint. We first would like to distinguish water *consumption* from *withdrawal*. Water withdrawal refers to getting water from the source (e.g., underground, rivers, sea) [35], and water consumption refers to "losing" water (e.g., by evaporation) in the process of withdrawal and return. Thus, water consumption is the *net* difference between water withdrawal and return. In this paper, we focus on water consumption, which is consistent with the industry standard [32]. The water consumption in data centers has two parts: *on-site direct* water and *off-site indirect* water.

On-site direct water consumption. Nearly all the server energy is converted into heat that must be removed from the data center server room to avoid overheating. Despite recent advances in cooling solutions, cooling towers are dominantly the most common cooling solution for warehouse-scale data centers, even for some leading companies such as Google [17] and Microsoft [26], and consume a huge amount of water. As illustrated in Figure 3, there are two water loops: one closed loop between the chiller and data center server room, and one open loop between the cooling tower and the chiller. Within the closed loop, water is not lost — it is pumped from the chiller into the data center to cool down the air handling unit's supply air in order to maintain a proper server inlet temperature, and warm water that absorbs the server heat returns to the chiller direction. Through a heat exchanger at the chiller, the heat is transferred from the closed loop to the open loop. Note that the chiller may operate in a "bypass" mode for energy saving when the outside temperature is low. Along the open loop, some of the water gets evaporated (i.e., "consumed") in the cooling tower to dissipate heat into the environment. Additionally, there is a process called "blown down" that drains the cooling water to reduce salt concentration accumulated in the cooling tower and hence also consumes water: the higher water quality, the more cycles of concentrations (i.e., water recirculates more times before "blown down") and hence the less blown-down water [36]. Through evaporation in the cooling tower as well as "blow down", data centers consume a significant amount of freshwater. For example, by one estimate [31] and depending on climate conditions, roughly 3.8 liters of water are consumed for each kWh of cooling load (approximately equal to server energy) by an average data center, resulting in a water usage effectiveness (WUE) of 3.8 L/kWh, while some data centers can even use 5.2 L/kWh [37]. Importantly, the on-site water comes from clean freshwater sources in order to avoid corrosion, clogged water pipes, bacterial growth, etc. While air-side economizers (a.k.a. "free" outside air cooling) do not need cooling towers and hence can reduce on-site water footprint, they often have strict requirements on the climate condition and may not be suitable for all locations. Additionally, freshwater is needed for humidity control to ensure proper server operation in data centers [38].

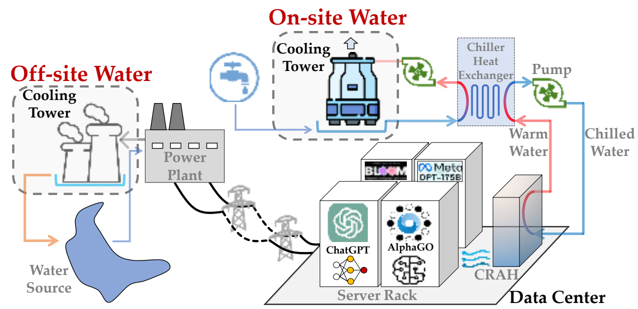


Figure 3: *Data center water footprint: on-site water consumption for data center cooling, and off-site water consumption for electricity generation. Multiple AI models are trained and/or deployed in the data center. The icons for AI models are only for illustration purposes.*

Off-site indirect water consumption. Data centers are held accountable for carbon footprint because of their (non-renewable) electricity usage. Likewise, electricity generation requires a huge amount of water [35, 39], thus resulting in off-site indirect water consumption for data centers [32]. Depending on the cooling technologies employed, different power generation systems (e.g., nuclear, coal and natural gas) consume different amounts of water. Among the non-hydroelectric systems, nuclear power plants typically consume the highest amount of water for producing each kWh of electricity on average, followed by thermoelectricity [35]. The water efficiency metric is called Energy Water Intensity Factor (EWIF) with a unit of L/kWh. Although certain renewable energy generation such as wind turbine and solar PV panels consumes no water for operation, their percentages in the overall electricity generation is still low in most parts of the world. Overall, excluding hydroelectricity (which itself does not consume water for cooling but expedites the evaporation of down streams), the national average EWIF by aggregating different types of power plants in the U.S. is about 1.8 L/kWh [32, 39].

3 Estimating Water Footprint of AI Models

While an AI model's water footprint depends in part on its energy consumption, such dependency is time-varying. As a result, simply multiplying the AI model's energy consumption by a constant and fixed WUE (Water Usage Effectiveness) does not yield an accurate estimate of AI models' water footprint. Next, by accounting for the time-varying WUE, we present a methodology for a fine-grained estimate of an AI model's water footprint.

3.1 Methodology

To obtain an AI model's total water footprint, we consider both on-site WUE and off-site WUE.

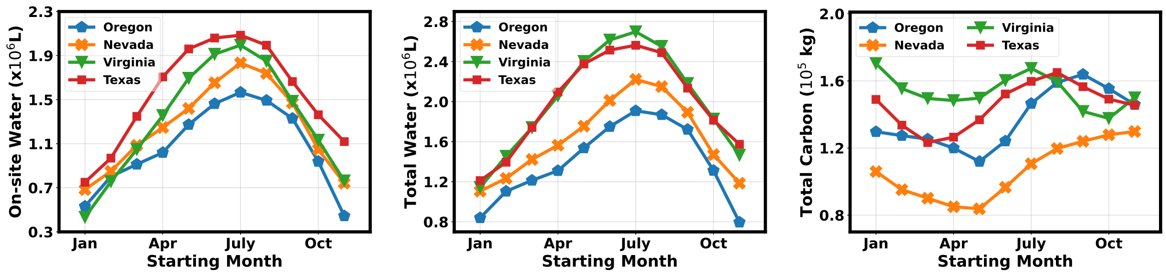
• **On-site WUE.** Cooling towers are most commonly used as the heat rejection mechanism for data centers. In general, the on-site WUE of cooling towers depends on multiple factors, such as temperature approach settings (i.e., difference between the cold water temperature and entering wet bulb temperature), cycles of concentrations (i.e., water recirculation times before "blown down"), water flow rate, air pressure, humidity, wet bulb temperature, and wind speed, among many others. Due to the lack of operational data from major data centers, we focus on the impact of outside wet bulb temperature on the on-site WUE, and present an empirical model based on a commercial cooling tower [40]. Specifically, following recommended operational settings, the on-site WUE can be approximated as $WUE_{on} = \frac{S}{S-1}\left(6 \times 10^{-5} \cdot T_{\mathrm{w}}^3 - 0.01 \cdot T_{\mathrm{w}}^2 + 0.61 \cdot T_{\mathrm{w}} - 10.40\right)$, where S is the cycle of concentrations and T_{w} is the outside wet bulb temperature (in Fahrenheit) [15]. The key insight for this formula is that the direct on-site WUE increases with outside wet bulb temperature, as a lower wet bulb temperature makes the water cool down more by the outside air and hence less through evaporation. When T_{w} is sufficiently low, the approximate formula for WUE_{on} can become negative and does not hold — water evaporation may not be needed for cooling in this case, and instead, water is mostly used for humidity control. Additionally, the empirical formula does not apply if the data center uses outside air cooling (e.g., in some of Meta's data centers [27]), and in this case, detailed operational data is needed to construct an accurate model for the on-site WUE.

• **Off-site WUE.** We now present the off-site indirect WUE measured in terms of EWIF (Electricity Water Intensity Factor). The same way as AI models are accountable for carbon footprint associated with off-site electricity generation, the off-site water footprint should also be taken into account to provide a more comprehensive environmental footprint [32]. Specifically, the off-site WUE depends on the energy fuel mixes (e.g., coal, nuclear, hydro) as well as cooling techniques used by power plants [32,41]. Since electricity produced by different energy fuels becomes non-differentiated once entering the grid, we consider the average EWIF, which can be estimated as $WUE_{off} = \frac{\sum_k b_k \times EWIF_k}{\sum_k b_k}$ where b_k denotes the amount of electricity generated from fuel type k for the grid serving the data center under consideration, and $EWIF_k$ is the EWIF for fuel type k [42,43]. As a result, variations in energy fuel mixes of electricity generation (to meet various demand levels) result in temporal variations of the off-site WUE. Moreover, the off-site WUE also varies by location, because each fuel type has its own distinct WUE [32] and energy fuel mix is typically different between states as some states may use less water-efficient energy generation than others [15,32,44].

• **Water footprint.** The on-site direct water consumption can be obtained by multiplying AI's energy consumption with the on-site WUE, while the indirect water consumption depends on the electricity usage as well as the local off-site WUE. Consider a time-slotted model $t = 1, 2, \cdots, T$, where each time slot can be 10 minutes to an hour depending on how frequently we want to assess the water footprint, and T is the total length of interest (e.g., training stage, total inference stage, or a combination of both). At time t, suppose that an AI model uses energy e_t (which can be measured using power meters and/or servers' built-in tools), the on-site WUE is $WUE_{on,t}$, the off-site WUE is $WUE_{off,t}$, and the data center hosting the AI model has a power usage effectiveness (PUE) of PUE_t that accounts for the non-IT energy such as cooling systems and power distribution losses. Then, the total water footprint W of the AI model can be written as

$$W = W_{on} + W_{off} = \sum_{t=1}^{T} e_t \cdot WUE_{on,t} + \sum_{t=1}^{T} e_t \cdot PUE_t \cdot WUE_{off,t}, \tag{1}$$

where $W_{on} = \sum_{t=1}^{T} e_t \cdot WUE_{on,t}$ and $W_{off} = \sum_{t=1}^{T} e_t \cdot PUE_t \cdot WUE_{off,t}$ are the on-site and off-site water footprints, respectively.



(a) On-site Water Footprint (b) Total Water Footprint (c) Total Carbon Footprint

Figure 4: *Estimated water and carbon footprints of training LaMDA with different starting months in 2022.*

Table 1: *Estimated EWIF for Common Energy Fuel Types in the U.S.* [45].

Fuel Type	Coal	Nuclear	Natural Gas	Solar (PV)	Wind	Other	Hydro
EWIF (L/kWh)	1.7	2.3	1.1	0	0	1.8	68 (0, if excluded)

Our methodology for estimating AI models' water footprint is general and applies to data centers with any type of cooling systems. For example, if the data center uses a cooling tower other than the one we model, we only need a different $WUE_{on,t}$. Naturally, given more operational data and transparency, both $WUE_{on,t}$ and $WUE_{off,t}$ can be refined to yield more accurate estimates.

3.2 Example: Water Footprint of LaMDA

We now use Google's large language model — LaMDA (Language Models for Dialog Applications) [34] — as an example and estimate its water footprint using our methodology. LaMDA uses about 451 MWh for training, but the starting time, specific data center locations, runtime PUE, WUE and EWIF for training and deploying LaMDA are not publicly disclosed. Thus, to estimate the water footprint of LaMDA, we use our empirical model presented in Section 3.1 as well as the state-level EWIF, and consider four different locations of Google's U.S. data centers — Loudoun County (VA), Henderson (NV), Midlothian (TX), and The Dalles (OR). As a result, our estimated water footprint of LaMDA only serves as an approximate reference point for the research community and general public, rather than an accurate calculation that is impossible to know without further transparency from the model developer.

More concretely, we use the hourly weather data [46] and state-level energy fuel mix data [47] for the year of 2022 in each of the chosen data center locations. We obtain the wet bulb temperature from the dry bulb temperature and relative humidity based on [48], and use the EWIF data shown in Table 1 for each energy fuel type. As the EWIF for each energy fuel type depends on the cooling techniques, the EWIF data we use is calculated as the weighted average of the median EWIF in [45] each cooling technique, excluding hydropower. We set the PUE as 1.1 (which is representative of Google's best data center efficiency). Based on the hourly wet bulb temperature and setting the cycle of concentration as $S = 5$, we use our empirical formula in Section 3.1 to derive the on-site WUE, whose minimum is capped at 0.01L/kWh.

With a total training time of consecutive 57.7 days [34], we consider different starting dates and data center locations for training LaMDA, and show the resulting on-site and total water footprints in Figure 4. While it is impossible to know the actual water footprint without detailed information from Google, our estimate shows that the total water footprint of training LaMDA is in the order of million liters. Crucially, we see that different training months and data center locations can significantly affect the water footprint — in general, summer may *not* be a good time for training large AI models due to excessive on-site water evaporation. Our empirical estimate shows that the highest total water footprint can be more than 3 times the lowest, highlighting the potential of cutting water footprint by judiciously scheduling AI model training.

For comparison, we also show in Figure 4 the estimated carbon footprint for training LaMDA. Note that [34] uses a fixed carbon efficiency of 0.056kg/kWh to calculate the carbon footprint of LaMDA without disclosing further details (e.g., time-varying energy fuel mix). Here, to estimate the carbon footprint of LaMDA, we use the state-level average EWIF to calculate the off-site carbon efficiency by following the approach considered in [43].[4] While our estimated carbon footprint differs from the value reported by [34], the key message we would like to highlight is that the most carbon-efficient training months and data center location may not be water-efficient. For example, due to the high penetration of solar energy, the carbon footprint of training LaMDA in Nevada is low, but the resulting water footprint can be very high due to high temperatures. This suggests that carbon-efficient scheduling of AI model training may not enable truly sustainable AI.

4 Limitations of the Existing Approaches

We now review some of the existing approaches to achieving sustainable AI, and discuss their limitations.

Improving on-site water efficiency. The existing approaches to improving on-site water efficiency are mostly from the "engineering" perspective (e.g., improving the data center's cooling tower efficiency) [17, 26,27,49,50]. These water-saving approaches can be viewed as *supply*-side solutions — saving water while supplying enough cooling to meet the given demand. But, the *demand*-side management — cooling demands are affected by "when" and "where" AI models are trained and used — is not addressed. While air-side economizers (a.k.a. "free" outside air cooling) can reduce on-site water footprint, they may have strict requirements on climate conditions and still use water for humidity control [27]. For example, even tech giants such as Google heavily rely on cooling towers and consume billions of liters of on-site cooling water each year [17]. Additionally, all these approaches only focus on the on-site water footprint, whereas the off-site water footprint that has time-varying off-site WUE due to variations in energy fuel mixes is not addressed.

Carbon-aware approaches. The existing research on sustainable AI has been primarily focused on AI models' carbon footprint [8–11,28,29,34]. Nonetheless, despite the correlation between water footprint and carbon footprint, the existing techniques for carbon efficiency do not necessarily equate to optimal water efficiency. This is because WUE varies with time and location in its own unique manner subject to real-time weather conditions and the power grid's mix of energy fuel sources. For example, AI model developers may want to train their models during the noon time when solar energy is more abundant [10], but this is also the hottest time of the day that leads to the worst water efficiency [44]. As a result, carbon-efficient scheduling of AI training can even lead to an increase in the water footprint, and reconciling such water-carbon conflicts requires new and holistic approaches to enable truly sustainable AI.

5 Our Findings and Recommendations

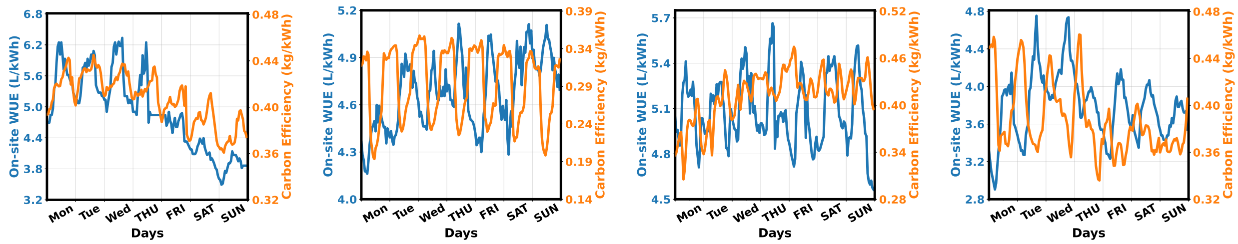
We provide our findings and recommendations to address the water footprint of AI models, making future AI more socially responsible and environmentally sustainable.

5.1 "When" and "Where" Matter!

Figure 4 shows that "when" and "where" to train a large AI model can significantly affect the water footprint. The underlying reason is the *spatial-temporal* diversity of both on-site and off-site WUE — on-site WUE changes due to variations of outside weather conditions, and off-site WUE changes due to variations of the grid's energy fuel mixes to meet time-varying demands. In fact, WUE varies at a much faster timescale than monthly or seasonably. For example, we show hourly on-site WUE and total WUE for the first week of August 2022 in Figure 5 and Figure 6, respectively. We see that both on-site WUE and total WUE vary significantly over time, and they are not *synchronized* across different locations. Therefore, by exploiting *spatial-temporal* diversity of WUE, we can dynamically schedule AI model training and inference to cut the water footprint. For example, if we train a small AI model, we can schedule the training task at midnight and/or in a data center location with better water efficiency. Likewise, some water-conscious users may prefer to use the inference services of AI models during water-efficient hours and/or in water-efficient data centers, which can contribute to the reduction of AI models' lifecycle water footprint. Such *demand*-side water management, which decides "when" and "where" to train and use AI models by exploiting the spatial-temporal WUE, complements the existing engineering-based on-site water saving approaches that focus on the supply side.

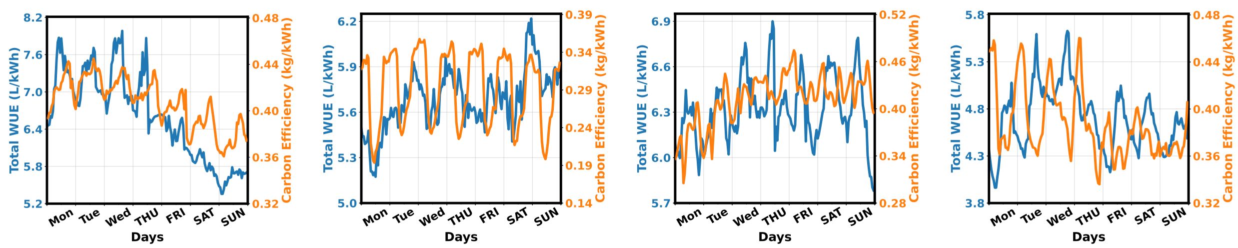
[4] Since electricity generated by different fuel types cannot be differentiated once entering the grid, [43] uses the weighted average carbon efficiency of different fuel types serving a data center for carbon footprint estimation. This approach essentially considers the actual carbon emission, without accounting for carbon footprint reduction through purchased carbon credits.

As a side note, federated learning [51], where multiple users collaboratively train an AI model using their own datasets on their local devices (e.g., personal computers) without consuming on-site water for data center cooling, can also leverage the spatial-temporal diversity of off-site WUE to reduce the total water footprint. Specifically, by integrating water-efficient schedules into local AI model training, we can achieve water footprint reduction for federated learning. Similar practices have been adopted for carbon footprint reduction in real products — Apple has recently integrated clean energy scheduling into its iPhone by selecting low-carbon hours for charging [52], and Microsoft launched carbon-aware Windows Update services by scheduling installations at specific times of the day when the off-site carbon efficiency is better [53].



(a) Virginia (b) Nevada (c) Texas (d) Oregon

__Figure 5:__ Hourly carbon efficiency and on-site WUE for the first week of August 2022.



(a) Virginia (b) Nevada (c) Texas (d) Oregon

__Figure 6:__ Hourly carbon efficiency and total WUE for the first week of August 2022.

5.2 More Transparency is Needed.

To exploit the spatial-temporal diversity of WUE, it is crucial to have a better visibility of the runtime WUE and increase transparency by keeping the AI model developers as well as end users informed of the runtime water efficiency.

Modeling the runtime WUE. The complex relations between the on-site WUE and multiple runtime factors (e.g., outside temperature, humidity, water flow rate, among others) would probably mandate a more advanced and sophisticated model than the one we have presented in Section 3.1. To capture the complex relations, we can leverage strong predictive power of a neural network parameterized by θ to estimate the on-site WUE as $WUE_{on} = f_\theta(x)$ where x represents the input features. Clearly, building such a neural network model requires labeled data provided by data center operators, but such data is still lacking in the public domain.

Increasing transparency. With more transparency, the general public can be better engaged into the global efforts to address the growing water challenges. For example, water-conscious users may prefer to use the inference services of AI models during water-efficient hours, but currently, the lack of transparency about AI models' detailed water efficiency prohibits them from doing so. Additionally, being informed of the data center's runtime water efficiency, AI model developers can better schedule their model training and choose locations for the deployment of trained models.

We recommend AI model developers and data center operators be more transparent. For example, what are the runtime (say, hourly) on-site WUE and off-site WUE? When and where are the AI models trained?

What about the AI models trained and/or deployed in third-party colocation data centers or public clouds? Such information will be of great value to the research community and the general public.

5.3 "Follow the Sun" or "Unfollow the Sun"?

To cut the carbon footprint, it is preferable to "follow the sun" when solar energy is more abundant. Nonetheless, to cut the water footprint, it is more appealing to "unfollow the sun" to avoid high-temperature hours of a day when (on-site) WUE is high. This conflict can also be reflected by Figure 5 and Figure 6 where we see that the carbon efficiency and WUE do not align very well with each other. Our figures (e.g., Figure 8) in the appendix also demonstrate this point: carbon-efficient hours and water-efficient hours are different. Thus, to judiciously achieve a balance between "follow the sun" for carbon efficiency and "unfollow the sun" for water efficiency, we need to reconcile the water-carbon conflicts by using new and holistic approaches. In other words, only focusing on AI models' carbon footprint alone is far from enough to enable truly sustainable AI.

6 Conclusion

In this paper, we recognize the enormous water footprint as a critical concern for socially responsible and environmentally sustainable AI, and make the first-of-its-kind efforts to uncover the secret water footprint of AI models. Specifically, we present a principled methodology to estimate the fine-grained water footprint, and show that AI models such as GPT-3 and Google's LaMDA can consume a stunning amount of water in the order of millions of liters. We also show that WUE is varying both spatially and temporally — judiciously deciding "when" and "where" to train a large AI model can significantly cut the water footprint. In addition, we point out the need of increasing transparency of AI models' water footprint, and highlight the necessity of holistically addressing water footprint along with carbon footprint to enable truly sustainable AI.

AI models' water footprint can no longer stay under the radar — water footprint must be addressed as a priority as part of the collective efforts to combat global water challenges.

References

[1] Facebook. Sustainability report, 2020, https://sustainability.fb.com/wp-content/uploads/2021/06/2020_FB_Sustainability-Report.pdf.

[2] Amazon. Sustainability — water stewardship, 2023, https://sustainability.aboutamazon.com/environment/the-cloud/water-stewardship.

[3] Google. Water commitments, 2023, https://sustainability.google/commitments/water/.

[4] UN Water Conference. How 'aquapreneurs' are innovating to solve the water crisis, 2023, https://www.weforum.org/agenda/2023/03/un-water-conference-aquapreneurs-innovation.

[5] U.S. White House. White House action plan on global water security, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/06/water-action-plan_final_formatted.pdf.

[6] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2), feb 2022.

[7] A. Shehabi, S. J. Smith, N. Horner, I. Azevedo, R. Brown, J. Koomey, E. Masanet, D. Sartor, M. Herrlin, and W. Lintner. United States data center energy usage report. *Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775*, 2016.

[8] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.

[9] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12):54–63, nov 2020.

[10] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics.

[11] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[12] The water resources group: Background, impact and the way forward, http://www3.weforum.org/docs/WEF/WRG_Background_Impact_and_Way_Forward.pdf.

[13] United Nations. Water scarcity, https://www.unwater.org/water-facts/water-scarcity.

[14] United Nations Children's Fund (UNICEF). Water scarcity — addressing the growing lack of available water to meet children's needs., https://www.unicef.org/wash/water-scarcity.

[15] Mohammad. A. Islam, Shaolei Ren, Gang Quan, Muhammad Z. Shakir, and Athanasios V. Vasilakos. Water-constrained geographic load balancing in data centers. *IEEE Trans. Cloud Computing*, 2015.

[16] Md Abu Bakar Siddik, Arman Shehabi, and Landon Marston. The environmental footprint of data centers in the United States. *Environmental Research Letters*, 16(6):064017, 2021.

[17] Urs Hölzle. Our commitment to climate-conscious data center cooling, 2022, https://blog.google/outreach-initiatives/sustainability/our-commitment-to-climate-conscious-data-center-cooling/.

[18] BMW Group. Next level mobility, 2022, https://www.bmwgroup.com/en/report/2022/downloads/BMW-Group-Report-2022-en.pdf.

[19] Tesla. Impact report, 2021, https://www.tesla.com/ns_videos/2021-tesla-impact-report.pdf.

[20] U.S. Drought Monitor. Map Archive. https://droughtmonitor.unl.edu/Maps/MapArchive.aspx.

[21] DatacenterMap. Colocation USA, http://www.datacentermap.com/usa/.

[22] Palm Beach Daily News. Drought dilemma: Palm Beach leaders struggle to cap water use, https://www.palmbeachpost.com/story/weather/2011/06/09/water-managers-shortage-near-crisis/7179447007/.

[23] Water constraints in Singapore, http://water-constraints.blogspot.com/.

[24] Myeongjae Jeon, Shivaram Venkataraman, Amar Phanishayee, unjie Qian, Wencong Xiao, and Fan Yang. Analysis of large-scale multi-tenant gpu clusters for dnn training workloads. In *Proceedings of the 2019 USENIX Conference on Usenix Annual Technical Conference*, USENIX ATC '19, page 947–960, USA, 2019. USENIX Association.

[25] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, and Xiaodong Wang. Applied machine learning at facebook: A datacenter infrastructure perspective. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 620–629, 2018.

[26] Melanie Nakagawa. The journey to water positive, 2023, https://blogs.microsoft.com/on-the-issues/2023/03/22/water-positive-climate-resilience-open-call/.

[27] Meta. Sustainability — water, https://sustainability.fb.com/water/.

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[29] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training, 2021.

[30] Noelle Walsh. How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability, 2022, https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability/.

[31] Drew FitzGerald. Data centers and hidden water use The Wall Street Journal, 2015, https://www.wsj.com/articles/data-centers-1435168386.

[32] The Green Grid. Water usage effectiveness (WUE): A green grid data center sustainability metric. *Whitepaper*, 2011.

[33] Kasper Groes Albin Ludvigsen. ChatGPT's electricity consumption, pt. ii, 2023, https://kaspergroesludvigsen.medium.com/chatgpts-electricity-consumption-pt-ii-225e7e43f22b.

[34] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language models for dialog applications, 2022.

[35] Jordan Macknick, Robin Newmark, Garvin Heath, and KC Hallett. A review of operational water consumption and withdrawal factors for electricity generating technologies. *NREL Tech. Report: NREL/TP-6A20-50900*, 2011.

[36] Environmental Defense Fund. How cooling towers work, `http://business.edf.org/projects/attwater/how-cooling-towers-work`.

[37] Minglei Zhang, Yunpeng Gao, and Zhengyong Wang. Analysis of annual water consumption for a data center in beijing, china, 2022, `https://blog.csdn.net/j6UL6lQ4vA97XlM/article/details/126112615`.

[38] D. Lee. Water efficiency at Facebook's Prineville data center, `https://www.opencompute.org/blog/water-efficiency-at-facebooks-prineville-data-center`.

[39] Paul Allen Torcellini, Nicholas Long, and Ron Judkoff. Consumptive water use for US power production, 2003.

[40] SPX Cooling. Water usage calculator, `http://spxcooling.com/green/leed/water-usage-calculator/`.

[41] U.S. Dept. of Energy. Energy demands on water resources, Dec. 2006.

[42] Kishwar Ahmed, Mohammad A. Islam, Shaolei Ren, and Gang Quan. Exploiting temporal diversity of water efficiency to make data center less 'thirsty'. In *ICAC*, 2014.

[43] Peter Xiang Gao, Andrew R. Curtis, Bernard Wong, and Srinivasan Keshav. It's not easy being green. *SIGCOMM Comput. Commun. Rev.*, 2012.

[44] 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.

[45] Jordan Macknick, Robin Newmark, Garvin Heath, and KC Hallett. A review of operational water consumption and withdrawal factors for electricity generating technologies. *NREL Tech. Report: NREL/TP-6A20-50900*, 2011.

[46] Iowa State University. Iowa environmental mesonet, `https://mesonet.agron.iastate.edu/`.

[47] U.S. Energy Information Administratio. Open data, `https://www.eia.gov/opendata/`.

[48] D. Meyer and D. Thevenard. PsychroLib: A library of psychrometric functions to calculate thermodynamic properties of air. *Journal of Open Source Software*, 4(33):1137, 2019.

[49] AT&T Water Management, `https://about.att.com/csr/home/reporting/issue-brief/water-management.html`.

[50] U.S. EPA. Watersense, `http://www3.epa.gov/watersense/commercial/`.

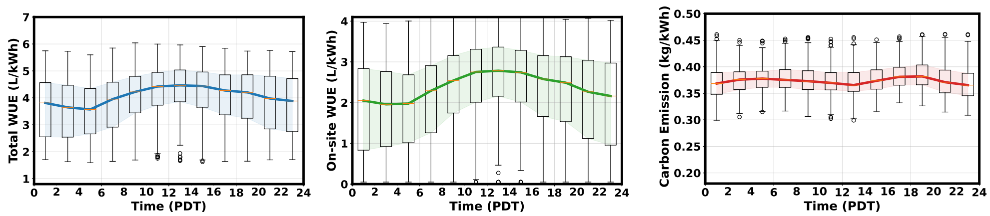
[51] Jakub Konecny, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[52] Use Clean Energy Charging on your iPhone. Apple, `https://support.apple.com/en-us/HT213323`.

[53] Microsoft. Windows update is now carbon aware, `https://support.microsoft.com/en-us/windows/windows-update-is-now-carbon-aware-a53f39bc-5531-4bb1-9e78-db38d7a6df20`.

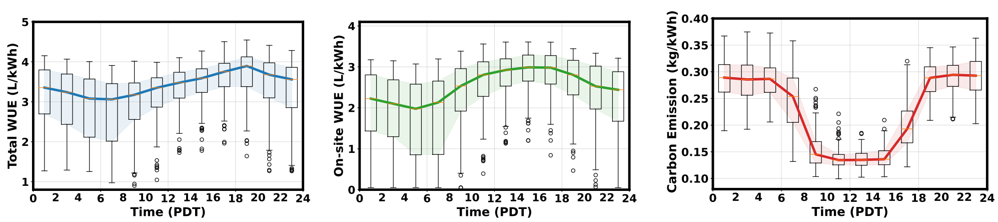
Appendix

We present the average hourly WUE (both on-site and total) as well as hourly carbon efficiency over different months for different data center locations. These figures demonstrate that the hourly WUE can vary significantly over time and across different locations. Importantly, WUE and carbon efficiency are not necessarily compatible with each other, calling for new research to reconcile the conflicts and enable truly socially responsible and environmentally sustainable AI.



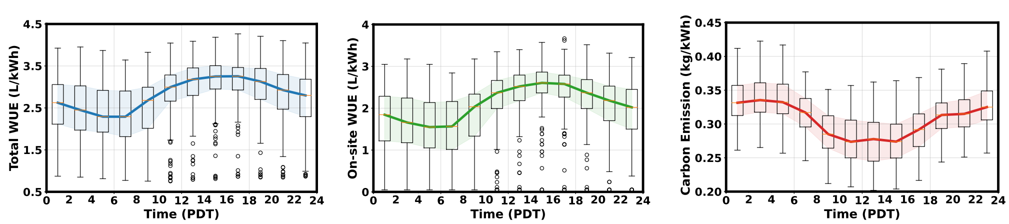
(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 7: *Average hourly WUE and carbon efficiency in Loudoun County, Virginia. February 1 to April 30, 2022.*



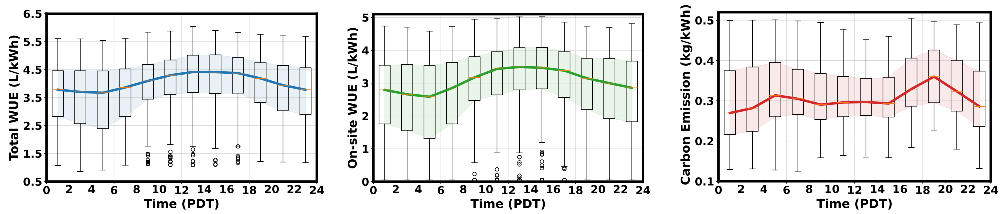
(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 8: *Average hourly WUE and carbon efficiency in Henderson, Nevada. February 1 to April 30, 2022.*



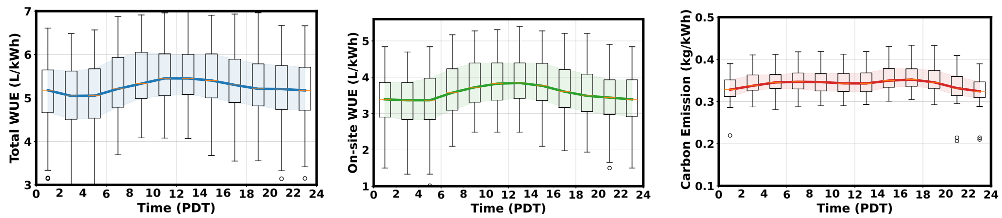
(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 9: *Average hourly WUE and carbon efficiency in The Dalles, Oregon. February 1 to April 30, 2022.*



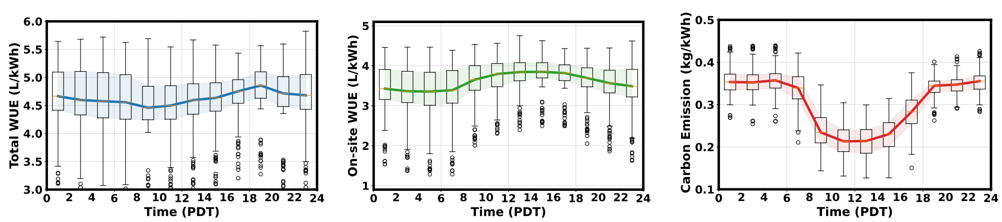
(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 10: *Average hourly WUE and carbon efficiency in Midlothian, Texas. February 1 to April 30, 2022.*



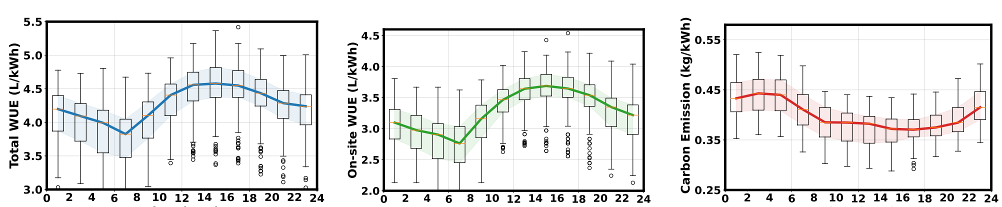
(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 11: *Average hourly WUE and carbon efficiency in Loudoun County, Virginia. September 1 to October 30, 2022.*



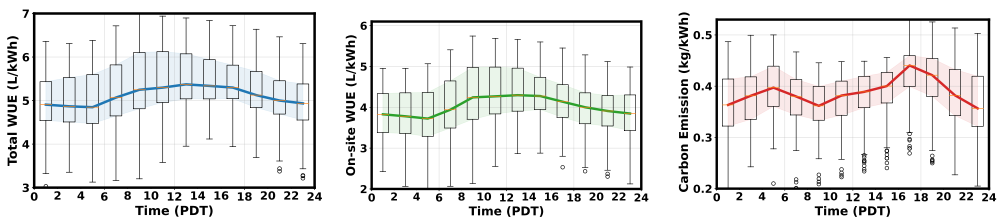
(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 12: *Average hourly WUE and carbon efficiency in Henderson, Nevada. September 1 to October 30, 2022.*



(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 13: *Average hourly WUE and carbon efficiency in The Dalles, Oregon. September 1 to October 30, 2022.*



(a) Total WUE (b) On-site WUE (c) Carbon Efficiency

Figure 14: *Average hourly WUE and carbon efficiency in Midlothian, Texas. September 1 to October 30, 2022.*