

Employing federated learning for training autonomous HVAC systems

Fredrik Hagström^a, Vikas Garg^{b,c}, Fabricio Oliveira^{id a,*}

^a Department of Mathematics and Systems Analysis, Aalto University, Espoo, Finland

^b Department of Computer Science, Aalto University, Espoo, Finland

^c YaiYai Ltd, Espoo, Finland

ARTICLE INFO

Keywords:

Federated learning
Reinforcement learning
HVAC control
Energy consumption
Thermal comfort
Soft actor-critic

ABSTRACT

Buildings account for 40 % of global energy consumption. A considerable portion of building energy consumption stems from heating, ventilation, and air conditioning (HVAC), and thus implementing smart, energy-efficient HVAC systems has the potential to significantly impact the course of climate change. In recent years, model-free reinforcement learning algorithms have been increasingly assessed for this purpose due to their ability to learn and adapt purely from experience. They have been shown to outperform classical controllers in terms of energy cost and consumption, as well as thermal comfort. However, their weakness lies in their relatively poor data efficiency, requiring long periods of training to reach acceptable policies, making them inapplicable to real-world controllers directly.

In this paper, we demonstrate that using federated learning to train the reinforcement learning controller of HVAC systems can improve the learning speed, as well as improve their ability to generalize, which in turn facilitates transfer learning to unseen building environments. In our setting, a global control policy is learned by aggregating local policies trained on multiple data centers located in different climate zones. The goal of the policy is to simultaneously minimize energy consumption and maximize thermal comfort. We perform a thorough set of experiments, evaluating three different optimizers for local policy training, as well as three different federated learning algorithms against two alternative baselines. We demonstrate through experimental evaluation that these effects lead to a faster learning speed, as well as greater generalization capabilities in the federated policy compared to any individually trained policy. Furthermore, the learning stability is significantly improved, with the learning process and performance of the federated policy being less sensitive to the choice of parameters and the inherent randomness of reinforcement learning.

1. Introduction

One of the greater challenges of modern society is that of climate change. Efforts to mitigate climate change must focus not only on the supply side of energy, e.g., renewable and nuclear energy, but also on the demand side, considering factors such as energy consumption and efficiency [1]. Of the global energy consumption, buildings alone are responsible for roughly 40 % of the total consumption [2]. Heating, ventilation and air conditioning (HVAC) are major factors in building energy consumption [1], and hence, developing smart and energy-efficient HVAC control systems can play an important role in mitigating climate change.

Most of the current HVAC systems in residential buildings are managed by classical algorithms, such as rule-based controllers and proportional, integral and derivative controllers [2]. These controllers not only lack knowledge of the thermal dynamics of the building environment but are also unable to take weather predictions into account. Hence,

they are unable to react and adapt to changes in the environment, leading to sub-optimal energy performance [3]. To utilize predictive data and knowledge of the building environment for improved building control performance, one can rely on Model Predictive Control (MPC) techniques [3]. MPC can anticipate when to, e.g., preheat a building based on weather and occupant forecasts, in order to improve energy efficiency. MPC has been shown to be effective at reducing energy consumption on both simulated and real building environments [3]. However, a serious drawback to MPC is that it requires accurate models of the environment in which the controller operates. On the other hand, every building is unique and, as such, developing a general MPC-based energy management system that can be deployed to various buildings is extremely difficult, and MPC is yet to be adopted by the building industry on a wider scale [3].

In recent years, through the emergence and rapid development of deep learning, it has become increasingly popular to apply machine learning techniques in multiple different research fields [4].

* Corresponding author.

E-mail address: fabricio.oliveira@aalto.fi (F. Oliveira).

<https://doi.org/10.1016/j.enbuild.2025.115761>

Received 18 November 2024; Received in revised form 15 April 2025; Accepted 16 April 2025

Available online 24 April 2025

0378-7788/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Reinforcement learning, a sub-field of machine learning concerned with control problems, has also started to gain considerable interest in research on energy system applications, including HVAC control systems. In particular, model-free reinforcement learning algorithms provide a promising direction for building control. As the name suggests, these algorithms do not require any model of the building environment or of its dynamics within. Instead, they learn purely from data collected while interacting with the environment. This eliminates the need for expert domain knowledge to develop models of the environment and allows the algorithms to be applied to any building in general, which are the main challenges of developing and deploying MPC-based controllers. Reinforcement learning algorithms also have greater adaptability to changes in the environment, as they can learn from the environment indefinitely, and as such, they can take into account long-term changes, such as changes in climate and occupant behaviour.

Reinforcement learning has been successfully applied for building-related control tasks, though mostly in simulated environments [3]. A major hurdle for the deployment of reinforcement learning algorithms to real buildings is their poor data efficiency. They need to collect large amounts of experience data to learn decision policies that take reasonable actions, and thus, it takes a long time to train them. For example, soft actor-critic, a state-of-the-art algorithm with best-performant data efficiency and learning speed requires more than a year of training to produce an acceptable policy in terms of thermal comfort [2]. Currently, this data inefficiency makes it infeasible to train reinforcement learning algorithms directly in physical building environments. A promising approach to overcome this data efficiency is to use transfer learning, i.e., to pre-train a controller on a simulated environment and then move it to a real environment for fine tuning [3]. Still, it is not known how to generalize a controller trained on a small set of buildings for use in another building not seen during training [3].

Our main contribution is to address this gap by investigating how federated learning can improve data efficiency and generalization in reinforcement learning-based HVAC control. To the best of our knowledge, this is the first study to systematically evaluate the impact of federated optimization on the learning dynamics and performance of reinforcement learning agents in a distributed HVAC control setting.

Federated learning is a paradigm for decentralized distributed machine learning [5]. A shared global model is trained on data distributed locally over a network of participating nodes by sending copies of the global model to the nodes, training the copies on the local data, and sending the local updates back to a central server for model aggregation. The local data of each node is never explicitly shared with other nodes, nor with the central server. This reduces the communication costs associated with transmitting data and eliminates the need for large storage capacity at the coordinating central server, while simultaneously ensuring a higher degree of data privacy at the nodes. Federated learning also makes no assumptions about the distribution of the data, and thus it can be applied to systems with heterogeneous components. These features make federated learning an ideal distributed learning scheme for smart HVAC system controllers, since every building will have its own unique data distribution, and sensitive information, e.g., occupancy behaviour, will be kept private. By training a controller on multiple buildings simultaneously, we effectively collect the total experience data at a higher rate than any single building, which counteracts the low data efficiency of reinforcement learning algorithms. Also, since the data distribution is heterogeneous, the collected experience data varies from building to building, leading the total experience to be more diverse, thereby facilitating greater generalization capabilities in the shared controller.

In this paper, we demonstrate the effectiveness of federated learning for training reinforcement learning-based HVAC controllers. In a real-world deployment, the proposed system would consist of a network of HVAC controllers operating across multiple buildings, each equipped with local reinforcement learning agents. These agents would collect sensor and forecasted data (e.g., temperature, air relative humidity, and energy consumption; for a complete list of those used in our experi-

ments, please see Table B.4) and update their control policies accordingly. Instead of training in isolation, the agents would participate in a federated learning framework, where local updates are periodically aggregated at a central server to refine a global control policy. This global policy is then redistributed to each building, aiming to improve learning efficiency and enable generalization across different environments.

We perform an experimental evaluation of a federated controller trained in multiple simulated data center environments using the *Federated Averaging* algorithm [5]. The objective of the controller is to minimize energy consumption while maintaining thermal comfort, i.e., keeping the temperature within a user-specified (deemed acceptable) range of values. We evaluate and compare the performance of three different optimizers on the local nodes: *stochastic gradient descent*, *stochastic gradient descent with momentum*, and *Adam* [6]. The performance of the federated controller with the best local optimizer is then compared to that of individual controllers trained exclusively on each respective data center. Furthermore, we evaluate two additional federated learning algorithms: *Federated Averaging with server momentum* [7] and *FedAdam* [8]. We apply a *gradient masking* technique [9] to each federated algorithm to improve learning stability. The reinforcement learning algorithm used is the Soft Actor-Critic (SAC) algorithm [10–12], which has previously shown to outperform other alternatives in HVAC control tasks [2,13]. Our main findings from applying federated learning to train a reinforcement learning HVAC control agent are:

- Improved generalization: The federated control agent outperforms all individual agents when applied to an unseen environment.
- Increased learning speed: The federated control agent is able to converge to the best policy at a faster rate than an individually trained agent.
- Improved learning stability: There is an inherent randomness to the training process of reinforcement learning agents. Federated learning reduces the variance across different training runs, leading to more consistent results.
- Benefits of adaptivity: The federated training process can benefit from adaptivity on the local optimizers, as Adam outperforms stochastic gradient descent with and without momentum.

The rest of this paper consists of the following parts. Section 2 presents an overview of the related literature. Section 3 discusses the methodology used in this paper by presenting the key technical aspects related to the SAC algorithm, as well as those related to federated learning. Section 4 describes the simulated environment and the setup of our experiments, as well as the results obtained. Finally, we draw conclusions in Section 6.

2. Related work

In this section, we provide a non-exhaustive survey of the literature on the employment of reinforcement learning for HVAC control. The purpose is to provide a general overview of different techniques available, whilst delineating our contribution to the literature. For more extensive surveys, we refer the reader to the works of Wang and Hong [3], Perera and Kamalaruban [4], Vázquez-Canteli and Nagy [14] and Weinberg et al. [15].

2.1. Early approaches

Some of the first applications of reinforcement learning to the control of HVAC systems were made around the turn of the millennium. Anderson et al. [16] combined a proportional plus integral (PI) controller with a reinforcement learning component to control a heating coil. They evaluated it in a simulated environment, showing improved performance compared to the PI controller alone. Mozer [17] utilized reinforcement learning in the control of the HVAC, Domestic Hot Water (DHW) and lighting systems of a real house, with the objective of minimizing both

electricity cost and occupant discomfort. In [18], the authors investigated a reinforcement learning solution for the operation of a simulated thermal storage system to reduce energy costs, showing favorable results when compared to conventional controllers. Liu and Henze [19] used Q-learning to train both passive and active thermal storage controllers for reduced energy costs. They found the performance to be sensitive to the learning parameters and the sizes of the state and action spaces. The training time was also observed to be unacceptably long for real-world applications. They followed up their research with a hybrid learning approach in Liu and Henze [20,21], where the agent is first pre-trained in a simulation of the environment, after which it is applied to and further trained on the true environment, making it an early example of transfer learning. They found the approach to significantly reduce the training time needed in the true environment. However, this approach requires an accurate model of the environment for the simulation phase, therefore having the same drawback as MPC.

2.2. Value-based approaches

The last decade has seen an increase in research on reinforcement learning in the energy domain, Perera and Kamalaruban [4], Vázquez-Canteli and Nagy [14]. Sun et al. [22] minimized the day-ahead energy costs using an event-based approach, where the reinforcement learning agent takes actions only “as needed”, instead of in regular time intervals. This reduces computational requirements while maintaining similar performance in cost savings and human comfort compared to time-based approaches. Barrett and Linder [23] reduced the cost of energy while meeting the temperature set-point specified by the user during periods of occupancy. They employ a Bayesian learning approach to predict occupancy and a Q-learning agent to control the thermostat unit. Li and Xia [24] trained a Q-learning agent to simultaneously minimize energy consumption and maximize thermal comfort. They improve upon the learning speed of standard Q-learning by utilizing a multi-grid approach, where the discretization of the state and action spaces are highly coarse at the beginning for early convergence, after which both spaces are iteratively refined during training for more fine control of the HVAC system. Ruelens et al. [25] minimized the energy costs of thermostatically controlled loads in both a dynamic pricing and day-ahead scheduling scenario using a Fitted Q-iteration controller equipped with a backup controller to ensure comfort. The controller converges much faster than standard Q-learning, and yields significant cost savings compared to the default controller, though increasing the energy consumption. A similar approach was taken by Costanzo et al. [26]. Wei et al. [27] minimized the energy costs and thermal comfort violations of a multi-zone building using a Deep Q-network (DQN), which achieves comparable levels of comfort violations while yielding greater cost savings than standard Q-learning.

The papers reviewed thus far focus on value-based reinforcement learning, in most cases Q-learning. Their limitations lie in that they must discretize the state and actions spaces, and scale poorly in terms of computation and memory to both the increase in dimension of the space and the granularity of the features [28,29]. Hence, in practice, the discretization is often coarse. HVAC control tasks are often naturally formulated as continuous control problems. Modeling the control task with value-based methods can therefore lead to oversimplification, since the rough discretization of state and action spaces sacrifices finer control. In contrast, policy gradient and actor-critic algorithms learn continuous policy functions and, as such, can provide more suitable alternatives.

2.3. Policy-based approaches

Policy gradient and actor-critic algorithms, while applicable to continuous control, have not seen nearly as much interest as value-based methods in the HVAC control literature, as well as building control in general [3,14]. This is likely due to earlier algorithms being either difficult to train due to high hyperparameter sensitivity or having poor data

efficiency, making them unfeasible for any potential real-world application [2].

Still, policy-based and actor-critic methods have been the algorithm of choice in some applications. Gao et al. [30] combined a deep neural network for thermal comfort prediction with Deep Deterministic Policy Gradient (DDPG) to control an HVAC system. DDPG is shown to outperform the value-based Q-learning, SARSA and DQN algorithms in terms of energy consumption and thermal comfort. A similar comparison and conclusion was made between DDPG and DQN in Du et al. [31]. Biemann et al. [2] evaluated and compared the performance of four actor-critic algorithms; Trust Region Policy Optimisation (TRPO), Proximal Policy Optimisation (PPO), Twin Delayed DDPG (TD3) and SAC, which have received little attention in the energy domain, despite their success in other domains [4]. Biemann et al. [2] concluded that while all four algorithms reduce energy consumption compared to their model-based baseline controller, SAC provides the best trade-off between energy savings and thermal comfort, while simultaneously displaying significantly greater learning speed and stability. In [32], PPO was used to reduce energy consumption while maintaining thermal comfort. The control policy was pre-trained on historical data of the existing controller using imitation learning. Thus, the policy learns to emulate the existing controller, performing reasonably well already at deployment, and quickly improving through fine-tuning with the PPO algorithm. The performance was evaluated in both simulated environments and a real conference room. The pre-trained PPO controller managed to reduce the cooling demand in the real environment, making the approach reasonable for real-world deployment, assuming the existence of historical controller data.

2.4. Model-based approaches

Model-based reinforcement learning approaches have also been explored, albeit the role the model plays varies. For example, in Gao and Wang [33], a model of the environment was learned through function approximation. The learned model is used to generate additional simulated experience in conjunction with the real experience, leading to faster convergence of the reinforcement learning algorithm. Nagy et al. [34] also learned a model of the environment, but instead used the model to plan the actions multiple steps ahead. While model-based approaches demonstrate greater sample efficiency than model-free algorithms, leading them to learn significantly faster, their success depends on how accurate the model is. In [34], their model-based algorithm converges in only about 20 days, while simultaneously outperforming the model-free approach in terms of both consumption and comfort. However, they showed that if the learned model is incorrect or if the dynamics of the environment change, the algorithm fails to adapt and is in turn outperformed by the model-free algorithm. As with MPC-based approaches, the main drawback of model-based approaches is that they require accurate models to achieve successful performance. As the dynamics of different buildings vary greatly and are difficult to model, developing model-based control systems that can be deployed generally is a challenging task.

2.5. Federated learning in the building domain

Federated learning has seen some application in the building energy domain. Khalil et al. [35] used Federated Averaging to train a thermal comfort predictive model, which is used as input for a rule-based temperature set-point controller. They follow up in Khalil et al. [36] with a modified implementation of Federated Averaging for reduced overhead in communication. Guo et al. [37] used federated learning to train machine learning models to predict the coefficient of performance of a chiller. Gao et al. [38] trained a federated model for forecasting the energy demand of buildings. Lu et al. [39] also take a federated approach to residential energy consumption forecasting, incorporating a reinforcement learning agent to assign weights to each local model when

performing model aggregation. In [40], federated learning was used to train a model for regulation capacity evaluation of an HVAC system. Lee et al. [41] used a federated reinforcement learning model to schedule the energy consumption of the HVAC systems of three buildings with solar photovoltaic systems and a shared controllable energy storage system. In [42], a similar approach to ours was taken, training a SAC agent for HVAC control using Federated Averaging, though in a notably different setting. They evaluate two different scenarios. In their power-saving scenario, the temperature setting of the AC is fixed, and the task of the agent is to turn the AC on when people are present in a room and off when the room is empty. In the second, normal operation scenario, the agent also aims to control the charging and discharging of a storage battery, with the goal of maintaining the temperature below a threshold. The agent is able to perform in the power-saving scenario, and Fujita et al. [42] observe an increase in the rate of convergence when using federated learning, but the agent is unable to achieve ideal control in the normal operation scenario.

Our survey suggests that, while there has been effort dedicated to the employment of reinforcement learning for controlling HVAC systems with a degree of success, there is a lack of focus on investigating whether federated learning can be used to address some of the challenges faced by these studies, such as data efficiency and generalization. This is precisely where our contribution lies. To the best of our knowledge, our work is the first to thoroughly evaluate the effects federated optimization has on the learning and performance of reinforcement learning agents for direct control of HVAC systems.

3. Methodology

3.1. Reinforcement learning

Reinforcement learning is, in its essence, a computational paradigm where how to optimize a decision-making problem is “learned by doing” [43]. The two main components of reinforcement learning are the agent and the environment. The agent aims to learn how to optimally interact with the environment in which it exists through trial and error. The agent-environment interaction follows a *Markov Decision Process* (MDP), which is a stochastic control process that evolves in a sequence of discrete time steps $t \in \mathbb{N}$. An MDP can be formally represented as a tuple (S, \mathcal{A}, R, P) , where

- S is the *state space*, i.e., the set of possible states s ,
- \mathcal{A} is the *action space*, i.e., the set of possible actions a ,
- $R : S \times \mathcal{A} \rightarrow \mathbb{R}$ is the *reward function* $R(s_t, a_t)$,
- $P : S \times S \times \mathcal{A} \rightarrow [0, 1]$ is the *transition probability function* $P(s' | s, a)$.

At time step t , the agent chooses and performs an action a_t based on the current state s_t . The environment then transitions to state s_{t+1} following the dynamics of the environment described by the transition probability function $P(s_{t+1} | s_t, a_t)$. As a consequence of its actions, the agent receives a reward $r_t = R(s_t, a_t)$, which measures the quality of the chosen action. This process continues in the same way, resulting in a sequence of states and actions:

$$(s_0, a_0, s_1, a_1, s_2, a_2, \dots).$$

This sequence is known as a *trajectory*. In the literature, it is also commonly referred to as an episode or a rollout.

To decide what action to take in state s_t , the agent follows a so-called *policy* π . The policy can be either deterministic or stochastic. A deterministic policy is defined as a mapping $\pi : S \rightarrow \mathcal{A}$, such that $a_t = \pi(s_t)$. A stochastic policy is a probabilistic function $\pi : \mathcal{A} \times S \rightarrow [0, 1]$, where $a_t \sim \pi(\cdot | s_t)$ and $\sum_{a_t \in \mathcal{A}} \pi(a_t | s_t) = 1$.

3.1.1. Soft actor-critic

Soft Actor-Critic (SAC) [10–12] is a state-of-the-art deep reinforcement learning algorithm that learns a continuous stochastic policy. It is *model-free*, meaning that the policy is learned without knowledge of

the transition dynamics P . It is also *off-policy*, meaning that it can learn from experience samples generated by any arbitrary policy, making it more sample-efficient than *on-policy* algorithms, which can only utilize samples collected from the current policy. These factors make SAC a suitable option for HVAC control, where environment dynamics are difficult to model, collecting experience is time expensive, and continuous actions allow for finer control.

The objective in classical reinforcement learning is to find the policy π that maximizes the expected *return*, i.e., the expected sum of rewards $\sum_t \mathbb{E}_{(s_t, a_t) \sim p_\pi} [R(s_t, a_t)]$, where p_π refers to the state-action marginal of the trajectory distribution induced by π . The SAC algorithm considers instead an alternative maximum-entropy objective by adding an entropy term to the expectation as follows

$$J(\pi) = \max_{\pi} \sum_t \mathbb{E}_{(s_t, a_t) \sim p_\pi} [(R(s_t, a_t) - \alpha \log \pi(\cdot | s_t))], \quad (1)$$

where $\alpha \in [0, \infty)$ is the temperature variable that controls the trade-off between exploration (entropy) and exploitation (reward maximization).

Haarnoja et al. [10,11,12] derived the SAC algorithm from an algorithm called *Soft Policy Iteration* (SPI). SPI learns a policy by repeating two main steps: *policy evaluation* and *policy improvement*. The policy evaluation step evaluates the soft *action-value function* $Q : S \times \mathcal{A} \rightarrow \mathbb{R}$ of the current policy π , i.e., the expected return of starting in state s , taking action a , and adhering to the policy thereafter. The soft Q-value is evaluated by iteratively updating the soft Q-function until convergence according to the soft *Bellman equation*

$$Q_{k+1}(s_t, a_t) = R(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p_s} [V_k(s_{t+1})], \quad (2)$$

where $\gamma \in [0, 1]$ is the *discounting factor*, p_s is the state marginal of the trajectory distribution induced by π , and $V : S \rightarrow \mathbb{R}$ is the soft *state-value function*, i.e., the expected return starting from state s and following policy π thereafter. The state-value function V is given by

$$V_k(s_t) = \mathbb{E}_{a_t \sim \pi} [Q_k(s_t, a_t) - \alpha \log \pi(a_t | s_t)]. \quad (3)$$

In the policy improvement step, the policy is updated towards the exponential of the soft Q-function. In practice, it is preferable to have tractable policies, so the policy is restricted to a set of policies Π , which can be, e.g., a family of parameterized distributions. In the update, the new policy must therefore be projected onto the set Π . Haarnoja et al. [11] use information projection, and so the new policy is computed, for all states $s \in S$, according to

$$\pi_{new} = \arg \min_{\pi \in \Pi} D_{KL} \left(\pi(\cdot | s_t) \parallel \frac{\exp \left(\frac{1}{\alpha} Q^{old}(s_t, \cdot) \right)}{Z^{old}(s_t)} \right), \quad (4)$$

where $Z^{old}(s)$ is a partition function that normalizes the distribution and D_{KL} is the Kullback-Leibler divergence.

SPI is only applicable to discrete state and action spaces. To extend SPI to continuous spaces, Haarnoja et al. [11] introduce function approximators for the soft Q-function Q_θ and policy π_ϕ , and alternate between optimizing their parameterization via gradient descent (instead of performing their evaluations) and policy improvement steps, yielding the SAC algorithm. The SAC algorithm models the soft Q-function using a neural network. The policy π_ϕ is typically modeled as a Gaussian distribution, where the mean μ_ϕ and standard deviation σ_ϕ vectors are given by a neural network. The Q-function is updated via gradient descent, by minimizing a loss function based on the Bellman equations:

$$\mathcal{L}(\theta) = \frac{1}{|D|} \sum_{(s_t, a_t, r_t, s_{t+1}) \in D} \frac{1}{2} (Q_\theta(s_t, a_t) - y)^2, \quad (5)$$

where D is a mini-batch of experience examples (s_t, a_t, r_t, s_{t+1}) , and y is the target of the Q-network, derived from combining Eqs. (2) and (3):

$$y = r_t + \gamma (Q_\theta(s_{t+1}, \tilde{a}_{t+1}) - \alpha \log \pi(\tilde{a}_{t+1} | s_{t+1})). \quad (6)$$

Here, the next action \tilde{a}_{t+1} is sampled from the current policy $\tilde{a}' \sim \pi_\phi(\cdot | s_{t+1})$. The update utilizes a *target* Q-network parameterized by $\bar{\theta}$

to stabilise training. The target Q-network is obtained by Polyak averaging the Q-network weights with smoothing constant ρ over the course of training as

$$\bar{\theta} \leftarrow \rho\theta + (1 - \rho)\bar{\theta}. \quad (7)$$

The policy update can be computed by minimizing the expected KL-divergence in Eq. (4) via gradient descent

$$J(\phi) = \mathbb{E}_{s_t \sim D} \left[\mathbb{E}_{a_t \sim \pi_{\phi}(\cdot|s_t)} [\alpha \log \pi(a_t|s_t) - Q_{\theta}(s_t, a_t)] \right]. \quad (8)$$

Notice that the expression has been multiplied by α and the constant partition function Z is ignored since it does not affect the gradient. The performance $J(\phi)$ is an expectation over actions, which are dependent on the policy parameters ϕ , and so it is not possible to get an estimate of the gradient based on Eq. (8) directly. To get an expression for the gradient of the performance that can be estimated with samples, Haarnoja et al. [11] use the *reparameterization trick*. The policy is reparameterized by the transformation

$$\hat{a} = f_{\phi}(\epsilon_t, s_t) \quad (9)$$

where ϵ is some noise sampled from a fixed distribution. The transformation depends on the policy distribution used. For example, Haarnoja et al. [11] use a squashed Gaussian in practice to ensure that the action values are bounded, in which case the appropriate transformation is

$$f_{\phi}(\epsilon_t, s_t) = \tanh(\mu_{\phi}(s_t) + \sigma_{\phi}(s_t) \odot \epsilon_t), \epsilon_t \sim \mathcal{N}(0, 1). \quad (10)$$

With the transformation, the performance is then rewritten as

$$J(\phi) = \mathbb{E}_{s_t \sim D, \epsilon_t \sim \mathcal{N}} [\alpha \log \pi(f_{\phi}(\epsilon_t, s_t)|s_t) - Q_{\theta}(s_t, f_{\phi}(\epsilon_t, s_t))]. \quad (11)$$

One can notice that the expectation is no longer dependent on the policy parameters and so the gradient can be moved into the expectation and approximated. The full SAC algorithm is presented in Algorithm 2 in Appendix A.

3.2. Federated learning

Federated learning is a framework for learning a shared global model on decentralized data across multiple nodes, without the nodes sharing their private data. Unlike typical distributed learning, federated learning makes no assumptions about the data distribution across nodes being *independent and identically distributed* (IID), and so it can be applied to non-IID settings as well. Furthermore, federated learning can also handle unbalanced data, i.e., some nodes having significantly larger local data sets than others. These characteristics allow federated learning to take advantage of massive amounts of data spread out over a large, heterogeneous network, e.g., pictures taken and stored on mobile phones, to learn a global model that generalizes well, while never communicating the local data itself. This maintains a higher degree of privacy across nodes while simultaneously eliminating the need for a central data center capable of storing the entire global data set.

Federated learning is well-suited for smart HVAC system controllers due to its ability to accommodate the unique data distribution of each building and maintain the privacy of potentially sensitive information such as occupancy behavior. By training a controller across multiple buildings at once, we indirectly gather experience data more efficiently compared to training on a single building, which helps overcome the data efficiency limitations of reinforcement learning algorithms. Additionally, the heterogeneous data distribution results in more diverse experience data from different buildings, enhancing the generalization capabilities of the controller agent.

3.2.1. Federated averaging

The federated learning setting consists of two main components. Firstly, we have a set of K nodes, referred to as *clients*, which compute updates to a shared global model independently of each other by training on their local data. Secondly, we have a central server, which

coordinates the clients and updates the global model. One round of communication between the server and clients consists of the server sending the current global model parameters to a fraction $C \in (0, 1]$ of clients, chosen at random, the chosen clients computing their local updates, and finally sending their respective locally updated parameters to the server for model aggregation.

The federated optimization algorithm presented by McMahan et al. [5] can be applied to any problem with a finite-sum objective of the form

$$\min_{w \in \mathbb{R}^d} f(w) \quad \text{where} \quad f(w) \equiv \frac{1}{N} \sum_{i=1}^N f_i(w). \quad (12)$$

When applying federated optimization to, e.g., an actor-critic algorithm, we are optimizing two different objectives, where $f(w)$ corresponds to both $\mathcal{L}(\theta)$ and $J(\phi)$. Assuming the global data set is partitioned over K clients, where \mathcal{P}_k denotes the set of indexes of data points at client k , with $n_k = |\mathcal{P}_k|$, the objective can be rewritten as

$$f(w) \equiv \sum_{i=1}^N \frac{n_k}{N} F_k(w) \quad \text{where} \quad F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} f_i(w). \quad (13)$$

McMahan et al. [5] focus on the application of federated optimization to deep learning models, which are typically trained using some variant of stochastic gradient descent (SGD) to optimize their objective, the loss function. Hence, they use a federated version of SGD, called *FedSGD*, as a starting point for their developed federated optimization algorithm. For one round of FedSGD, with fixed learning rate η and fraction $C = 1$, each client k computes the average gradient on their local data $g_k = \nabla F_k(w_t)$, where w_t is the current global model. The local gradients are then aggregated at the central server and used to update the model according to

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t) \quad \text{where} \quad \nabla f(w_t) = \sum_{k=1}^K \frac{n_k}{N} g_k. \quad (14)$$

An equivalent update to (14) can be performed by taking one step of gradient descent on each local model $w_{t+1}^k \leftarrow w_t - \eta \nabla g_k$, and then aggregating the local model parameters via the following weighted average

$$w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{N} w_{t+1}^k. \quad (15)$$

Since the update (15) is just an average over the parameters of each local model, it is possible to perform multiple local steps of gradient descent $w^k \leftarrow w^k - \eta \nabla F_k(w^k)$ before averaging in order to increase the amount of computation per communication round. This is the core of the *Federated Averaging* (FedAvg) algorithm.

3.2.2. FedOpt

In FedAvg, the updated global model parameters w_{t+1} are computed by averaging the updated local parameters w_{t+1}^k according to Eq. (15). Alternatively, this update can be performed by computing the “pseudo-gradient” Δ_{t+1} , which is the average of differences between the local parameters and the current global model, $\Delta_{t+1}^k = w_{t+1}^k - w_t$, and adding it to the current parameters according to

$$w_{t+1} \leftarrow w_t + \Delta_{t+1} \quad \text{where} \quad \Delta_{t+1} = \sum_{k=1}^K \frac{n_k}{N} \Delta_{t+1}^k. \quad (16)$$

Through this formulation, the server update in FedAvg can be viewed as taking one gradient ascent step using the pseudo-gradient and a *global learning rate* $\eta_g = 1$. Reddi et al. [8] recognize the possibility of choosing other values of η_g . They also suggest the possible use of alternative server update rules based on the pseudo-gradient, as well as utilizing other optimizers than SGD on the client side. Combining these ideas, Reddi et al. [8] generalize FedAvg into a framework called *FedOpt*, presented in Algorithm 1.

Algorithm 1 FedOpt.

```

1: Initialise global model  $w_0$ 
2: for each communication round  $t = 0, 1, \dots, T$  do
3:    $m \leftarrow \max\{C \cdot K, 1\}$ 
4:    $S_t \leftarrow$  random set of  $m$  clients
5:    $w_{k,0}^t = w_t, \forall k \in S_t$ 
6:   for each client  $k \in S_t$  in parallel do
7:     for  $u = 0, 1, \dots, U - 1$  do
8:       Compute estimate  $g_{k,u}^t$  of  $\nabla F_k(w_{k,u}^t)$ 
9:        $w_{k,u+1}^t = \text{ClientOpt}(w_{k,u}^t, g_{k,u}^t, \eta_l, t)$ 
10:    end for
11:     $\Delta_t^k = w_{k,U}^t - w_t$ 
12:  end for
13:   $n_{tot} = \sum_{k \in S_t} n_k$ 
14:   $\Delta_t = \sum_{k \in S_t} \frac{n_k}{n_{tot}} \Delta_t^k$ 
15:   $w_{t+1} = \text{ServerOpt}(w_t, \Delta_t, \eta_g, t)$ 
16: end for

```

ClientOpt and ServerOpt in Algorithm 1 refer to the optimizers used at the clients and server, respectively. Any gradient-based optimizer can be applied. The hyperparameter η_l sets the *local learning rate* at the clients. The hyperparameter U determines how many local updates to perform in each communication round. Reddi et al. [8] also allow the optimizers to depend on the communication round t to facilitate the potential use of learning rate schedulers.

FedAvgM, which stands for Federated Averaging with Server Momentum [7], slightly modifies the FedAvg algorithm by adding a momentum term v . During a server update (line 15 in Algorithm 1), the momentum is updated according to

$$v_t \leftarrow \mu v_{t-1} + \eta_g \Delta_t, \quad (17)$$

where $\mu \in [0, 1)$ determines the level of momentum. The global weight parameters are then updated using the momentum as

$$w_{t+1} \leftarrow w_t + v_t. \quad (18)$$

FedAdam is an adaptation of the Adam optimizer [6] to ServerOpt, presented by Reddi et al. [8]. FedAdam uses two momentum terms m and v in the server update. The first momentum m is computed as the exponential moving average

$$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) \Delta_t \quad (19)$$

and the second momentum as the squared exponential moving average

$$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) \Delta_t^2, \quad (20)$$

where $\beta_1, \beta_2 \in [0, 1)$ are hyperparameters. The global model update is then computed according to

$$w_{t+1} \leftarrow w_t + \eta_g \frac{m_t}{\sqrt{v_t} + \epsilon}. \quad (21)$$

Here, $\epsilon > 0$ controls the *degree of adaptivity*.

3.2.3. Gradient masking

Gradient masking [9] can improve the performance of FL algorithms in heterogeneous settings. The idea of gradient masking is to apply a soft mask to the server update, which assigns higher importance to the components of the pseudo-gradients which are in agreement with the dominant direction, thereby better capturing the invariances across clients. In [13] gradient masking was found to improve the stability of the learning process by reducing the randomness across different seeds. The importance is determined by the sign agreement across parameters over the client updates Δ_t^k . Tenison et al. [9] define the agreement score $A \in [0, 1]$, which is given by

$$A \equiv \left| \frac{1}{K} \sum_{k=1}^K \text{sign}(\Delta_t^k) \right|. \quad (22)$$

The agreement score is then used to compute the mask \tilde{m}_t element-wise according to

$$[\tilde{m}_t]_j = 1 \quad \text{if} \quad A_j \geq \tau \quad \text{else} \quad A_j, \quad (23)$$

where $\tau \in (0, 1]$ is a hyperparameter determining the desired level of agreement. The mask \tilde{m}_t is then applied to the final computed update in ServerOpt before addition to the current model parameters via the element-wise product. The updates of FedAvg (16), FedAvgM (18) and FedAdam (21) with gradient masking are thus

$$\text{FedAvg: } w_{t+1} \leftarrow w_t + \tilde{m}_t \odot \Delta_{t+1} \quad (24)$$

$$\text{FedAvgM: } w_{t+1} \leftarrow w_t + \tilde{m}_t \odot v_t \quad (25)$$

$$\text{FedAdam: } w_{t+1} \leftarrow w_t + \eta_g \tilde{m}_t \odot \frac{m_t}{\sqrt{v_t} + \epsilon}. \quad (26)$$

4. Experiments

4.1. Simulation environment

In our experiments, we use the open-source building simulation and control framework Sinergym (v.2.0.0) [44]. Sinergym provides an interface for interacting with the building energy model simulation tool EnergyPlus via the OpenAI Gym API [45], a popular API for implementing and evaluating reinforcement learning algorithms. Sinergym provides a handful of different building environments as well as several weather profiles. We conduct our experiments on the available data center environment.¹ The data center has a total area of 491.3 m². It is split into two asymmetrical zones; the west and east zone, equipped with their own respective HVAC systems. The HVAC systems are composed of air economizers, evaporative coolers, a direct expansion cooling coil, a chilled water coil and a variable air volume fan. The heating and cooling set-points of each zone are controllable, and one episode of simulation runs for one year.

4.1.1. Markov decision process formulation

To apply reinforcement learning algorithms to the control of the HVAC systems, we must provide an MDP formulation of the building environment. We define a state space S , an action space \mathcal{A} and a reward function R . One environment step, or control action, is taken every 15 min within the simulation, leading to a total number of 35 040 steps for one simulation episode.

The agent observes a state vector $s \in S \subset \mathbb{R}^{18}$ of 18 features. The complete list of features is presented in Table B.4 in Appendix B. The features consist of the factors that we aim to control, namely the temperature of the zones and indirectly the energy consumption of the IT equipment and HVAC system, as well as other factors that relate to the temperature in the zones, e.g., outside air temperature. We also include “forecasted” outside temperature and air relative humidity values. This allows the agent to anticipate large changes in temperature and potentially counteract them by pre-heating or pre-cooling the zones. How the forecasted values are observed is described further in Appendix C.

The control variables of the data center model are the heating and cooling setpoint temperatures of each zone, and so the action $a \in \mathcal{A} \subset \mathbb{R}^4$ taken by the agent is a vector of 4 features, which determines these setpoint temperatures. The action space is described in Table 1. The actions are bounded by a range of possible values, which also include “bad” values that can lead the temperature in the zones to lie outside the comfortable range of values. The notion of good values should instead be encoded into the reward function and learned by the agent, irrespective of the possible range of values of the HVAC equipment available, as argued by Biemann et al. [2].

The goal is to train an agent that minimizes the total energy consumption of the data center. At the same time, the temperature inside

¹ The name of the environment file is 2ZoneDataCenterHVAC_wEconomizer.idf.

Table 1
Description of the action space.

Feature	Range	Unit
West zone cooling setpoint	[15.0, 22.5]	°C
West zone heating setpoint	[22.5, 30.0]	°C
East zone cooling setpoint	[15.0, 22.5]	°C
East zone heating setpoint	[22.5, 30.0]	°C

the building must remain within the target range. Hence, we need to encode information about the energy consumption and the *thermal comfort* into the reward signal. The reward function defined by Biemann et al. [2] does precisely this, and so, we use it in our MDP formulation. They define the following reward function

$$R(s) = r_{west} + r_{east} - \lambda_p(P_{it} + P_{hvac}), \quad (27)$$

where r_i is computed based on the thermal comfort in zone i , and P_{it} and P_{hvac} are the power demands of the IT and HVAC equipment, respectively. The term $\lambda_p \geq 0$ is a scaling factor for the energy component of the reward. Given the observed temperature T_i in zone i , the thermal comfort component is computed as

$$r_i = \exp(-\lambda_g(T_i - T_{tgt})^2) - \lambda_t(\max(T_{min} - T_i, 0) + \max(T_i - T_{max}, 0)), \quad (28)$$

where T_{tgt} is the desired target temperature, and T_{min} and T_{max} are the lower and upper bounds of the comfortable temperature range. Scalars $\lambda_g, \lambda_t \geq 0$ are hyperparameters that determine the shape of the reward function. The first term in Eq. (28) gives the function a Gaussian shape, with the purpose of motivating the agent to stay close to the target temperature, providing a more robust reward than a simple trapezoidal reward function. The second term, the trapezoid penalty, is added to extend the function to yield negative rewards far away from the center, helping the agent to better distinguish moderately bad actions from very bad ones than it would with the zero rewards of a simple Gaussian.

The thermal comfort reward r_i is close to 1 when the temperature of zone i is close to the target, and small or negative when close to or outside the comfort bounds. The total power demand $P_{tot} = P_{it} + P_{hvac}$ of the data center is in the order of 100 kW, and so to bring the energy penalty component in the reward function (27) to the same scale as the comfort component, we use the scaling factor $\lambda_p = 10^{-5}$ in our experiments. We set the comfort range bounds to $T_{min} = 18^\circ\text{C}$ and $T_{max} = 27^\circ\text{C}$ according to the recommended temperature range by the ASHRAE guidelines for data center power equipment [46]. The target temperature is set to the midpoint of the comfort range $T_{tgt} = (T_{min} + T_{max})/2$, so as to motivate the agent to stay as far away from the edges of comfort as possible. Finally, we set the hyperparameters $\lambda_g = 0.2$ and $\lambda_t = 0.1$ as in [2]. Fig. 1 displays the shape of the thermal comfort reward r_i with the chosen parameters.

Sinergym provides 12 different weather profiles from significantly different climates. Each profile is fixed and provides hourly weather observations over a one-year period. The training of our agents spans multiple years, and so we do not wish to use the same weather profile for every year of training since we cannot know if the agent learns a useful policy for variable weather or if it simply overfits the weather profile. Thankfully, Sinergym allows us to add stochasticity to the weather from year to year. In Appendix C we provide further details and the full list of the weather profiles considered in Table C.5.

4.2. Experiment configurations

We perform two main sets of experiments. In the first set, we train a federated HVAC control agent using FedAvg as the server optimizer. We evaluate the performance of three different client optimizers: SGD, SGD with momentum (SGDM), and Adam. We have 12 available weather profiles, and so we train on 11 client data centers, each with its own unique weather conditions. The Helsinki weather profile is reserved for

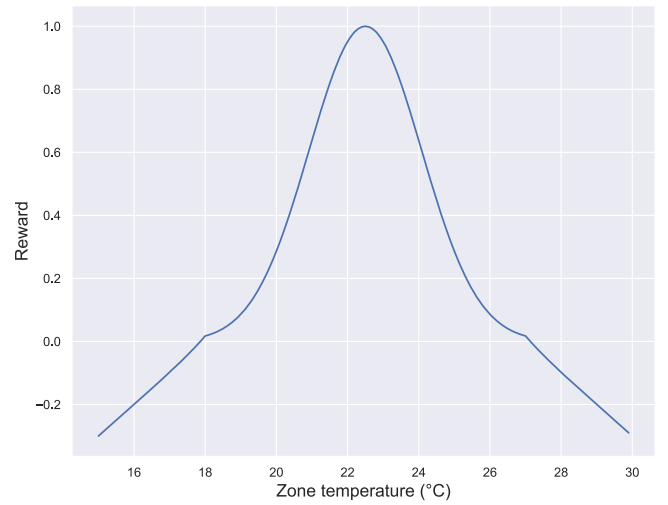


Fig. 1. Graph of the zone thermal comfort reward r_i . The hyperparameters are set to $\lambda_g = 0.2$ and $\lambda_t = 0.1$, and the comfort range is bounded to $T_{min} = 18^\circ\text{C}$ and $T_{max} = 27^\circ\text{C}$. The target temperature is set to the midpoint of the comfort range, $T_{tgt} = 22.5^\circ\text{C}$.

evaluating the performance of the global agent in unseen environments. We consider two performance comparison baselines. The first is the employment of a proportional-integral-derivative (PID) controller, using temperature as its process variable and defining its error according to the setpoints described in Fig. 1 and with hyperparameters set as described by [2]. This choice is justified by its widespread use in HVAC control applications. We also train individual agents for each client and include their performance as a baseline.

Lastly, in the second set of experiments, we evaluate two alternative federated algorithms, FedAvgM and FedAdam, using the best-performing client-side configuration from the first set.

Since our set of training clients is relatively small, we choose to include all clients in every global communication round, i.e., we set the fraction $C = 1$ for all our experiments. We also set the masking threshold to $\tau = 0.4$ in all experiments since it was found to generally perform well in Tenison et al. [9] and [13]. We evaluate FedAvg, and so the global learning rate is set to $\eta_g = 1$. For the client optimizers, we only vary the learning rate, and use the default values for other hyperparameters. See Table D.7 in Appendix D for the complete list of client optimizer hyperparameters. For the first set of experiments, we have two controllable hyperparameters, the client learning rate η_l and the total of local updates per round U . For each client optimizer, we perform a search over the following grid of values

$$\eta_l \in \{0.0003, 0.001, 0.01, 0.1\}$$

$$U \in \{4, 12, 24\}.$$

For FedAvgM, the controllable hyperparameters are the global learning rate η_g , the number of local updates per round U , and the server momentum β . We perform a search over the following grid of values

$$\eta_g \in \{0.001, 0.01, 0.1, 1.0\}$$

$$U \in \{4, 12, 24\}$$

$$\beta \in \{0.8, 0.9, 0.99\}.$$

For FedAdam we set the degree of adaptivity to $\epsilon = 10^{-3}$, as [8] find it to perform well across multiple different tasks. The controllable hyperparameters then are the global learning rate η_g , the number of local updates per round U , and the moment parameters β_1 and β_2 . We perform a search over the following grid of values

$$\eta_g \in \{0.001, 0.01, 0.1, 1.0\}$$

$$U \in \{4, 12, 24\}$$

Table 2

Performance of the federated agent for different client optimizers on the evaluation environment (Helsinki) after 15 episodes of training. We choose the configuration that yields the highest return for reporting the performance of the federated agent, which are $\eta_l = 0.001$, $U = 24$ for Adam, $\eta_l = 0.1$, $U = 24$ for SGD, and $\eta_l = 0.1$, $U = 12$ for SGDM. The reported values are the means over three episodes of evaluation. E_{tot} is the cumulative power consumption of the data center over one year, and Viol. is the comfort violation rate.

	E_{tot} (GWh)	Viol. (%)
Adam	0.9189	0.0016
SGDM	0.9220	0.0092
SGD	0.9266	0.0438
PID-Baseline	0.9311	0.0

$$\beta_1 \in \{0.8, 0.9, 0.99\}$$

$$\beta_2 \in \{0.9, 0.99, 0.999\}.$$

In all experiments, each configuration is repeated 3 times with different random seeds to evaluate the robustness of each configuration. The training runs over a period of 15 years. The simulator takes a step in the environment, i.e., sends observations to the agent and executes the actions chosen by the agent, every 15 min, and so a full training run consists of a total of 525,600 environment interactions. For further implementation details, see Appendix D. The source code is available at: <https://github.com/hagstromf/FedHVAC>.

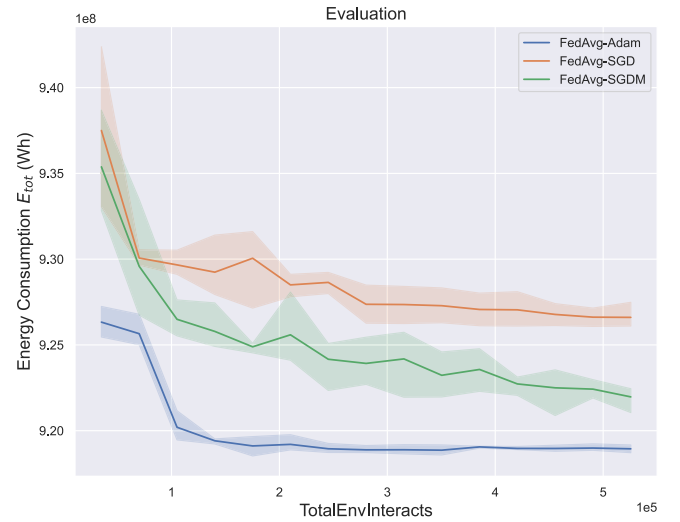
4.3. Results

In evaluating the performance of the models, we focus on the energy consumption and thermal comfort of the data center. The total energy consumption E_{tot} is the cumulative total power consumption P_{tot} over a year. The thermal comfort of the data center is evaluated in terms of thermal comfort violations. A thermal comfort violation takes place when the temperature in either or both zones of the building is outside the specified comfort range. The comfort violations are reported as the percentage of comfort-violating environment steps over a year.

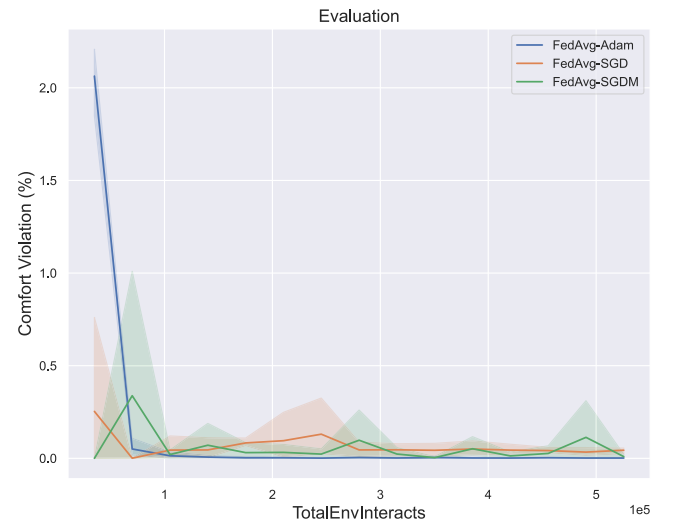
4.3.1. Evaluation results

First, we consider the performance of FedAvg using different client optimizers. At the end of training, each model is run for three episodes on the Helsinki evaluation environment. The results are presented in Table 2, where the performance values are the means over the three evaluation episodes over all three random seed iterations. We report the values of the configuration that yielded the highest mean return. Further discussion on the performance of different hyperparameter configurations is provided in Appendix E.

From Table 2, we notice that FedAvg with Adam outperforms SGD and SGDM in terms of both energy consumption and comfort violations when deployed on an unseen environment, indicating the best generalisation capabilities of the three. In Fig. 2, we show the progression of the energy consumption and comfort violation of the FedAvg agents on the evaluation environment for all client optimizers. The agents are evaluated for three episodes at the end of each episode of training. We plot the mean values over the three episodes over all random seeds with their bootstrapped 95 % confidence intervals. Based on the progression plots, we notice that FedAvg with Adam does not only offer improved generalisation compared to the others. It also displays faster learning speeds, with the energy consumption converging after about five episodes and the comfort violations converging after just three episodes, while SGD converges in roughly eight episodes, and SGDM has not converged yet at the end of training. FedAvg with Adam has better learning stability as well. The tighter confidence intervals regarding both energy consumption and comfort violation indicate that the learning is more robust to randomness in the model initialization and training process, and thus



(a) Energy consumption

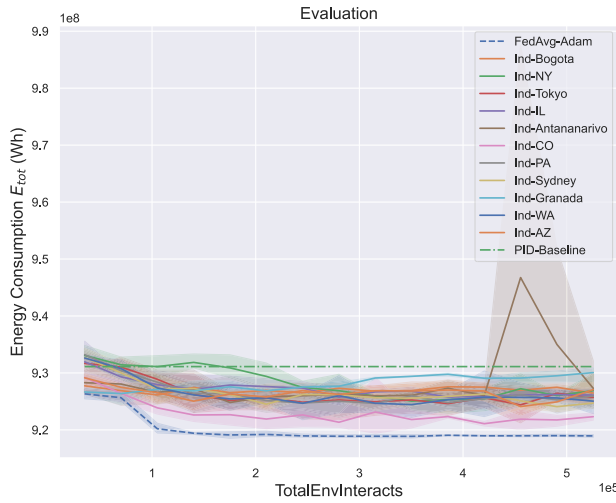


(b) Comfort violation

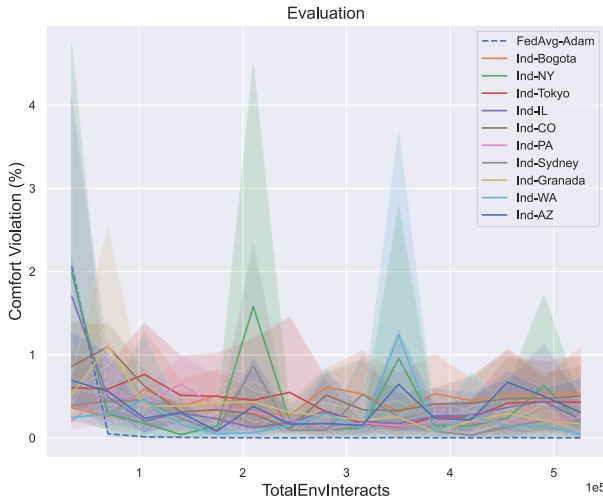
Fig. 2. Progression of the energy consumption and comfort violation on the Helsinki evaluation environment of the FedAvg agent with different client optimizers.

its performance is more reliable. Lastly, it can be noticed that all federated learning agents overperform the PID controller in terms of average power consumption. No comfort violation is observed for the PID controller.

Next, we analyze how the evaluation performance of a federated agent compares to agents trained independently on the clients. We focus on the best performing federated agent, i.e., with Adam as the client optimizer and $\eta_l = 0.001$, $U = 24$, and compare it to the best performing individual agents, with Adam and $\eta_l = 0.01$. The progression of energy consumption and comfort violations of the federated and independent agents in the evaluation environment are presented in Fig. 3. From Fig. 3a, we see that the federated agent outperforms every independent agent in terms of energy consumption, converging to a lower value, and at a faster rate. We also notice a high variance in the energy consumption, both across different clients as well as across different runs for each client, with a significant outlier in the agent trained in the Antanarivo environment. Remarkably, the variance for the federated agent is



(a) Energy consumption



(b) Comfort violation

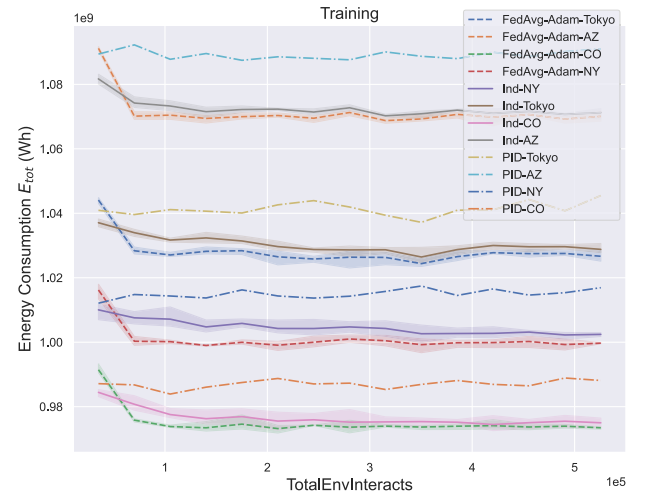
Fig. 3. Progression of the energy consumption and comfort violation on the Helsinki evaluation environment of FedAvg and independent agents with Adam as client optimizer. In the comfort violation plot 3b we omit the outlier Antananarivo for the sake of legibility.

significantly lower. Similar observations are made regarding the comfort violation in Fig. 3b, though we note that the independent agents tend to outperform the federated agent in the first episode.

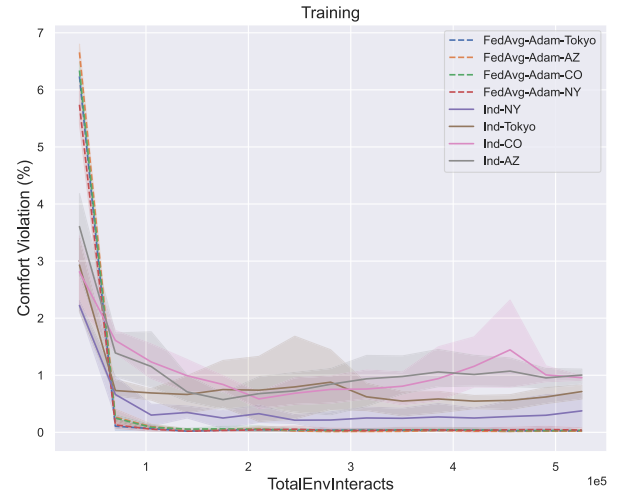
These observations support our conclusion that using federated optimization to train an HVAC control agent can significantly improve generalization, with a better performance in an unseen environment than any independently trained agent. Federated training can also improve the learning speed, generally converging faster, as well as learning stability, displaying a significant reduction in the variance in performance over different random seeds. In any real-world application, this consistency is a highly desirable trait, since we are not able to train the agent multiple times and therefore need a model that can reliably learn a good policy despite the inherent randomness of the real environment and training.

4.3.2. Training results

We have seen that applying federated optimization can improve the performance of a reinforcement learning HVAC control agent in an un-



(a) Energy consumption



(b) Comfort violation

Fig. 4. Progression of the energy consumption and comfort violation of FedAvg and independent agents on training environments Tokyo, AZ, CO and NY.

seen environment. Thankfully, this does not come at the expense of poorer performance in the training environments. In Fig. 4, we present the evolution of energy consumption and comfort violation of the federated agent and independent agents in the training environments. We only plot a subset of the environments for the sake of legibility. The behaviour of the omitted environments is consistent with the ones shown and analysed in this section. For additional figures of the remaining environments, please refer to Appendix F.

The energy consumption in Fig. 4a displays similar improvements from using FedAvg in the evaluation environment. FedAvg generally converges faster and manages to reach a lower level of energy consumption. We also see improved learning stability, with slightly less variance across training runs. These improvements are even more pronounced when analysing the progression of comfort violation in Fig. 4b. While the federated agent converges to near-zero comfort violations after two or three episodes, the independent agents never achieve near-zero violations. They also exhibit significantly more variance across both agents and different training runs. This shows that the model does not only benefit from the improved generalization, learning speed and learning stability of federated learning when applied to an unseen environment but also during training itself.

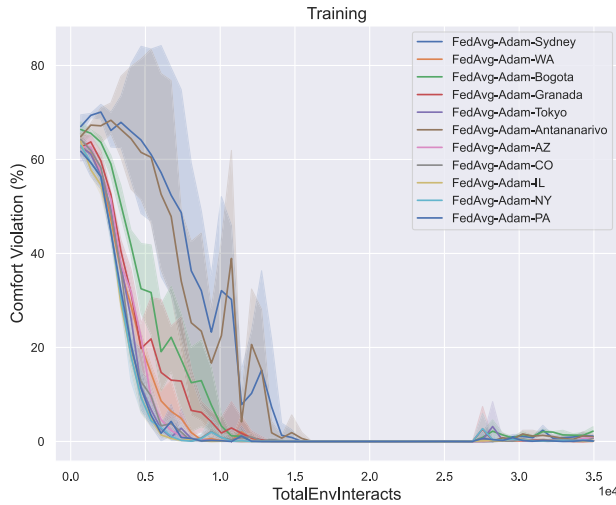
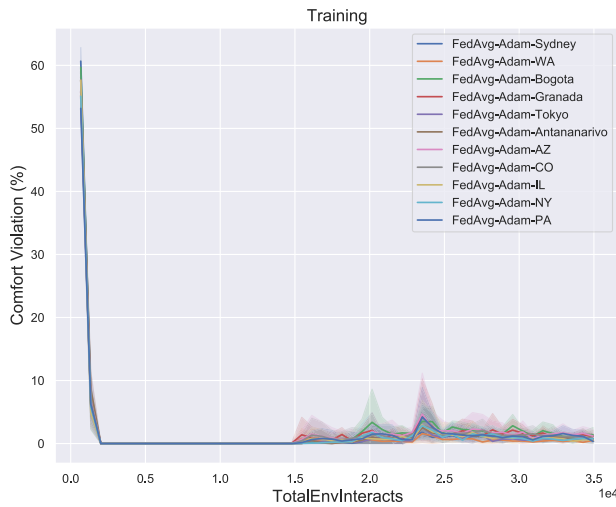
(a) $\eta_l = 0.001$ (b) $\eta_l = 0.01$

Fig. 5. Progression of the energy consumption and comfort violation of FedAvg and independent agents on training environments Tokyo, AZ, CO and NY.

While the federated agent outperforms the independent agents in the long run, we notice that the independent agents tend to perform better during the first episode, both in terms of energy consumption and comfort violation. This, however, seems to be an effect of the larger client learning rate η_l used for the independent agents. All federated agents perform better than the PID controller regarding energy consumption throughout the episodes.

In Fig. 5 we present the weekly comfort violations of the federated agent on the training environments over the first year of training for client learning rates $\eta_l = 0.001$ and $\eta_l = 0.01$. In this setting, the federated agent requires less than a full year of training to reach near-zero comfort violation. Depending on the environment, the agent with the lower client learning rate $\eta_l = 0.001$ requires between around 8000 to 17,000 steps to reach near-zero violations, corresponding to roughly 12 to 25 weeks. Some of the environments experience a small increase towards the end of the year. However, if we increase the client learning rate to $\eta_l = 0.01$, we can achieve near-zero comfort violation significantly faster, in just three weeks, though there is an increase in violations for the second half of the year. While increasing the client learning rate can

Table 3

Performance of the federated agent for different server optimizers on the evaluation environment (Helsinki) after 15 episodes of training. We choose the configuration that yields the highest return for reporting the performance of the federated agent, which are $\eta_g = 0.1$, $U = 24$, $\mu = 0.9$ for FedAvgM, and $\eta_g = 0.001$, $U = 24$, $\beta_1 = 0.8$, $\beta_2 = 0.9$ for FedAdam. The reported values are the means over three episodes of evaluation. E_{tot} is the cumulative power consumption of the data center over one year, and Viol. is the comfort violation rate.

	E_{tot} (GWh)	Viol. (%)
FedAvg	0.9189	0.0016
FedAvgM	0.9192	0.0035
FedAdam	0.9203	0.0092

lead to significantly faster comfort violation reduction, it comes at the cost of significantly worse performance in the long run (see Fig. E.10 in Appendix E.1). On the other hand, while training with a lower client learning rate leads to great performance, the violations during the first few months are inadmissible, and so this federated agent would not be suitable for a real-world setting.

4.3.3. Server optimizers

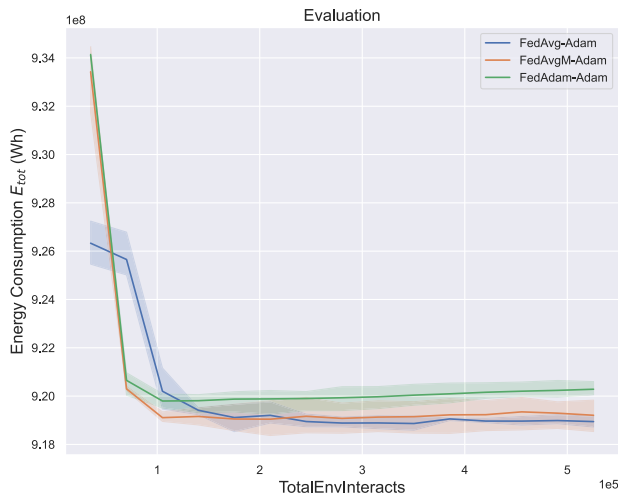
Besides FedAvg, we also evaluate two alternative server optimizers: FedAvgM and FedAdam. Both use Adam as the client optimizer, with $\eta_l = 0.001$. The evaluation performances of the best-performing configurations of each optimizer at the end of training are presented in Table 3. The progression plots of the energy consumption and comfort violations on the evaluation environment are presented in Fig. 6. From Table 3, we see that, although all perform similarly, FedAvg slightly outperforms the others in terms of both energy consumption and comfort violation. Looking at Fig. 6a and b, the most striking difference is the early performance of the optimizers. FedAvg has significantly worse comfort violations than FedAvgM and FedAdam in the first episode, but performs better in the second. The opposite is true for the energy consumption. We also notice that FedAvg has tighter confidence intervals, and so offers better learning stability than FedAvgM and FedAdam.

5. Discussion

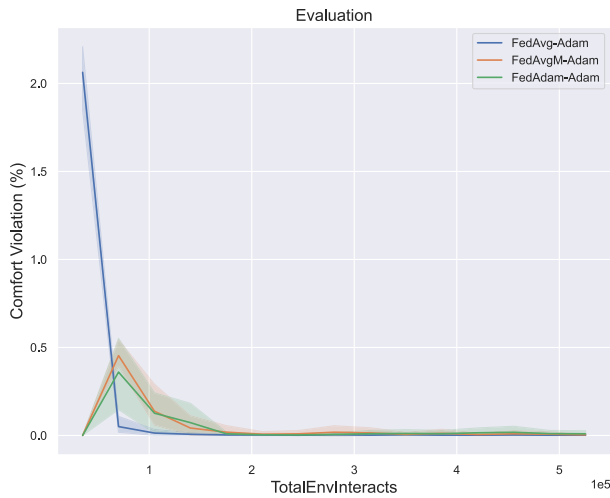
Through our experiments, we have identified three key improvements from applying federated optimization to training reinforcement learning HVAC controllers. Firstly, by learning from experience collected from multiple heterogeneous environments, the agent gains access, albeit indirectly, to a larger amount of training data, which generally encompasses more variability than that available to any independent agent. In other words, there is an increase in exploration, which leads to a more informed global agent that can *generalize better* to different environments. Secondly, the amount of total experience increases at a faster rate, which leads to an *increase in learning speed*. Finally, when aggregating over the local agents, the dominant direction in the pseudo-gradient will have the most impact on the global update. This seems to have a regularizing effect, making it more difficult for the agent to branch off into sub-optimal regions of the policy space, *increasing the learning stability*.

Our experiments show that the choice of client optimizer can have a significant impact on performance. The federated model can benefit from adaptivity on the local optimizer, as we found Adam to perform considerably better than both SGD and SGDM in terms of generalization, learning speed and learning stability. Meanwhile, the choice of server optimizer seems less critical.

Regarding the server optimizer, both FedAvgM and FedAdam display comparable performance to FedAvg, although FedAvg slightly outperforms them. Moreover, FedAvg also has the advantage of having fewer hyperparameters to be tuned. As shown in Appendix E.2, these algorithms can be considerably sensitive to the choice of said hyperparameters. In a real-world scenario, we cannot evaluate multiple different



(a) Energy consumption



(b) Comfort violation

Fig. 6. Progression of the energy consumption and comfort violation on the Helsinki evaluation environment of FedAvg, FedAvgM and FedAdam.

hyperparameters and, as such, it is desirable to use an algorithm with as few adjustable parameters as possible, with minimal sensitivity to said hyperparameters. Considering that FedAvg has fewer adjustable components, combined with the observation that neither FedAvgM nor FedAdam seems to offer any significant improvement in terms of either generalization, learning speed or learning stability, we believe that FedAvg provides a more defensible choice for future efforts related to deployment in real-world settings.

A few limitations of our experiments are worth highlighting. The training of the federated agent was revealed to be considerably sensitive to the clients' learning rate. With lower learning rates, federated optimization offers stable and fast learning, but it is not suitable as-is for a real-world building environment due to the high degree of comfort violations at the beginning of training. By increasing the learning rate, it is possible to significantly reduce the comfort violations early on, but this comes at a trade-off for significantly worse final performance. Second, although significant progress can be achieved by using federated learning in this particular context, challenges remain in bridging the gap between simulation and real-world deployment, which can be noted from the time taken, between 3 and 24 weeks depending on the

hyperparameter configuration, for the HVAC control agent to reach satisfactory energy consumption and comfort violation performance.

From a practical perspective in the context of HVAC control, federated learning offers significant benefits, being this a setting in which the underlying tasks across different buildings are largely similar yet subject to local variations. By allowing individual controllers to learn from their own operational data while sharing only aggregated model updates, the federated approach leverages commonalities across similar systems while preserving the confidentiality of sensitive information, such as occupancy patterns.

In our experiments, the primary focus was on energy consumption and comfort violations. However, practical real-world deployments would need to additionally account for communication or computational overhead. Nonetheless, the federated learning framework inherently reduces communication requirements by transmitting only aggregated model updates instead of raw data, while distributing the computational load across local nodes. This design suggests that, in a real-world HVAC system, the overhead from model aggregation is likely to be modest compared to the substantial benefits in terms of generalization, learning speed, and associated data privacy benefits.

6. Conclusion

In this paper, we have experimentally evaluated the effects of training reinforcement learning HVAC control agents via federated optimization. We have trained Soft Actor-Critic (SAC) agents using Federated Averaging (FedAvg) with gradient masking, evaluating and comparing the performance of three different client optimizers: stochastic gradient descent (SGD), stochastic gradient descent with momentum (SGDM), and Adam. We have also compared the performance of federated agents to that of individual agents, trained on each respective client environment used in the federated learning scenario, both in terms of their performance in an unseen test environment and their performance in the training environments themselves. Furthermore, two alternative server optimizers, Federated Averaging with server momentum (FedAvgM) and FedAdam were compared to the FedAvg algorithm.

Our results have demonstrated that federated learning can improve generalization and the learning speed and stability of reinforcement learning-based HVAC controllers, which are critical bottlenecks for their adoption in real-world settings. However, there are still important challenges that must be addressed in that direction, mainly related to the time required for the learning-based controllers to learn policies that perform satisfactorily.

Moreover, while our numerical experiments demonstrate clear benefits in terms of learning speed, generalization, and stability when employing federated optimization for reinforcement learning-based HVAC control, we acknowledge that these outcomes constitute only a first, albeit critical, step towards their wider deployment. As such, real-world pilot deployments remain essential to conclusively verify practical benefits and applicability in realistic building settings and thus warrant further research efforts.

Future research could be dedicated to bridging trade-offs between learning rates and comfort violation at the early stages of training through, e.g., the use of learning rate schedules, starting with a high learning rate and gradually decreasing it as training progresses. The great generalization of the federated agent provides another promising direction for future research. Practical implementations could benefit from integrating additional techniques in a hybrid manner, such as rule-based controllers (including the PID tested as baseline) or model-based approaches (if feasible) for improving early sample efficiency. Another promising direction is to focus on transfer learning from simulated to real environments, where a pre-trained agent is deployed and tuned on real buildings. Alternatively, the federated agent could also be pre-trained on historical data.

Finally, our choice of federated learning is driven by its pragmatic benefits—improving generalization, learning speed, and stability

through the aggregation of local updates while preserving data privacy. In contrast, meta-reinforcement learning (meta-RL), though promising for showing rapid adaptation between unseen tasks, still faces practical challenges, such as the need for meticulously curated task distributions and increased computational complexity. Nonetheless, as it develops further, meta-RL represents an interesting avenue for future research on autonomous HVAC control.

Data availability

Data and code will be made available in a github repository upon acceptance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Fredrik Hagström: Writing - review & editing, Writing - original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization; **Vikas Garg:** Writing - review & editing, Validation, Supervision, Methodology, Conceptualization; **Fabricio Oliveira:** Writing - review & editing, Writing - original draft, Supervision, Project administration, Funding acquisition, Conceptualization.

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project. This research received financial support from the Research Council of Finland (decision number 348092).

Appendix A. Soft actor-critic (SAC) pseudo-code

The final practical SAC algorithm used in our experiments includes a few additional features when compared to the algorithm presented in Section 3.1.1. The final algorithm learns two concurrent soft Q-functions, parameterized by $\theta_i, i \in \{1, 2\}$, which are trained independently to minimize $\mathcal{L}(\theta_i)$ in Eq. (5). They both have their respective target networks $\bar{\theta}_i, i \in \{1, 2\}$. The Eq. (6) for the target y is modified to utilize the minimum of the two Q-functions

$$y = r + \gamma \left(\min_{i=1,2} Q_{\bar{\theta}_i}(s', a') - \alpha \log \pi(a'|s') \right), \quad a' \sim \pi_\phi(\cdot|s') \quad (\text{A.1})$$

and similarly for the performance $J(\phi)$ in Eq. (11)

$$J(\phi) = \mathbb{E}_{s \sim D, \epsilon \sim \mathcal{N}} \left[\alpha \log \pi(f_\phi(\epsilon, s)|s) - \min_{i=1,2} Q_{\bar{\theta}_i}(s, f_\phi(\epsilon, s)) \right]. \quad (\text{A.2})$$

This double Q-learning trick is used to mitigate positive bias in the policy improvement step, which can degrade performance [11].

The SAC algorithm is particularly sensitive to the temperature α , which has to be fine-tuned to the task at hand in order to achieve appropriate performance. Haarnoja et al. [11] develop a method for automatically adjusting its value during training to stabilise learning across different tasks. The temperature is updated at each gradient step by minimizing the following objective

$$J(\alpha) = \mathbb{E}_{s \sim D, a \sim \pi_\phi} \left[-\alpha \log \pi_\phi(a|s) - \alpha \tilde{H} \right], \quad (\text{A.3})$$

where \tilde{H} is the minimum desired entropy. Haarnoja et al. [12] find that the algorithm is quite robust with respect to the minimum entropy, and generally setting it to -1 times the action dimension yields good results.

Algorithm 2 SAC.

```

1: Initialize:
   Critic networks  $Q_{\theta_1}, Q_{\theta_2}$  and actor network  $\pi_\phi$  with random parameters  $\theta_1, \theta_2, \phi$ .
   Target networks  $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2$ .
   Replay buffer  $\mathcal{B}$ .
2: for each iteration do
3:   for each environment step do
4:     Sample action  $a_t \sim \pi_\phi(\cdot|s_t)$  and observe reward  $r_t$  and next state  $s_{t+1}$ .
5:     Store transition tuple  $(s, a, r, s')$  in replay buffer  $\mathcal{B}$ .
6:   end for
7:   for each gradient step do
8:     Sample mini-batch  $\mathcal{D}$  from replay buffer  $\mathcal{B}$ .
9:     Compute targets  $y$  for all  $(s, a, r, s') \in \mathcal{D}$ , Eq. (A.1)
10:
11:     Update critics:  $\theta_i \leftarrow \theta_i - \lambda_Q \nabla_{\theta_i} \mathcal{L}(\theta_i)$ , Eq. (5).
12:     Update actor:  $\phi \leftarrow \phi - \lambda_\pi \nabla_\phi J(\phi)$ , Eq. (A.2).
13:     Update temperature:  $\alpha \leftarrow \alpha - \lambda_\alpha \nabla_\alpha J(\alpha)$ , Eq. (A.3).
14:
15:     Update target networks:  $\bar{\theta}_i \leftarrow \rho \theta_i + (1 - \rho) \bar{\theta}_i$ 
16:   end for
17: end for

```

Table B.4

Description of the state space.

Feature	Unit
Site Outdoor air drybulb temperature	°C
Site Outdoor Air Relative Humidity	%
Site Wind Speed	m/s
Site Wind Direction	degree
Site Diffuse Solar Radiation Rate per Area	W/m ²
Site Direct Solar Radiation Rate per Area	W/m ²
Zone Air Temperature(West Zone)	°C
Zone Air Relative Humidity(West Zone)	%
Zone Air Temperature(East Zone)	°C
Zone Air Relative Humidity(East Zone)	%
Facility Total HVAC Electricity Demand Rate	W
Facility Total Building Electricity Demand Rate	W
Forecasted Outdoor Air Drybulb Temp (+ 1 h)	°C
Forecasted Outdoor Air Relative Humidity (+ 1 h)	%
Forecasted Outdoor Air Drybulb Temp (+ 3 h)	°C
Forecasted Outdoor Air Relative Humidity (+ 3 h)	%
Forecasted Outdoor Air Drybulb Temp (+ 6 h)	°C
Forecasted Outdoor Air Relative Humidity (+ 6 h)	%

Appendix B. State space

Table B.4 shows the complete list of observed state features in the data center environment.

Appendix C. Weather profiles

To add stochasticity to the weather from year to year, Sinergym modifies the base weather profiles via the Ornstein-Uhlenbeck process at the beginning of each year. The Ornstein-Uhlenbeck process X_t is defined by the stochastic differential equation

$$dX_t = \tau(\mu - X_t)dt + \sigma dW_t, \quad (\text{C.1})$$

where W_t is Brownian motion with unit variance, and $\tau, \sigma \geq 0$ and μ are parameters affecting the evolution of the process. In our experiments, we set $\tau = 0.001$, $\sigma = 2.0$ and $\mu = 0$.

As mentioned in Section 4.1.1, we include “forecasted” outside temperature and relative humidity in our observations. These forecasted values are retrieved from the base weather profile. Since the weather over each year is stochastically modified from the base weather profile, the base weather profile provides us with values that are close to the

Table C.5

The base weather files available in Siner gym. M.T is the mean temperature and M.H is the mean relative humidity of the file.

Location	M.T (°C)	M.H (%)
Sydney, Australia	17.9	68.83
Bogota, Colombia	13.2	80.3
Granada, Spain	14.84	59.83
Helsinki, Finland	5.1	79.25
Tokyo, Japan	8.9	78.6
Antananarivo, Madagascar	18.35	75.91
Arizona, USA	21.7	34.9
Colorado, USA	9.95	55.25
Illinois, USA	9.92	70.3
New York, USA	12.6	68.5
Pennsylvania, USA	10.5	66.41
Washington, USA	9.3	81.1

“true” observed values, much like a typical weather forecast. Hence the base profile gives us a good proxy for a real weather forecast.

Appendix D. Implementation details

We use the implementation of the SAC algorithm provided by the Stable Baselines3 framework [47], which offers reliable implementations of reinforcement learning algorithms in PyTorch [48]. The Q-value functions and policy are approximated using simple feed-forward neural networks with an input layer, two hidden layers, and an output layer. As argued by Biemann et al. [2], in a real-world application, tuning all the hyperparameters of the algorithms becomes infeasible, hence the algorithms should perform well out-of-the-box. We therefore use the default hyperparameters of the Stable Baselines3 implementation. An exception is the rate at which the policy and Q-networks are updated. We set the training frequency to once every hour, i.e., after every 4 environment steps. At every update, the model takes a number of gradient steps equal to the number of environment steps taken between updates. See Table D.6 for a list of the exact hyperparameter values used for SAC.

In deep reinforcement learning and when training neural networks in general, it is often useful to ensure that all the features of the input vectors are on the same scale. This prevents very large features from dominating the calculated gradient, as well as maintains a more consistent range of values for the gradient, which often leads to more stable and faster learning. Hence we normalize the observations. The reward also affects the scale of the gradient, and as such, normalizing the rewards can also have a stabilizing effect. We therefore normalize the rewards as well. We use the VecNormalize wrapper in Stable Baselines3 with default values to normalize using a moving average.

SAC learns a stochastic policy. However, Haarnoja et al. [11] find that making the final policy deterministic often results in better performance than choosing actions stochastically, and so we set the SAC policy to be deterministic as well during evaluation. This is done by choosing the mean $\mu_\phi(s)$ of the policy distribution as the action.

Table D.6

SAC hyperparameters.

Critic networks	24 → 256 → 256 → 1
Actor networks	24 → 256 → 256 → (2 × 4)
Activation function	ReLU
Discount factor γ	0.99
Batch size	256
Polyak averaging ρ	0.005
Buffer size	10^6
Temperature α	auto
Target entropy	auto
Train frequency	4
Gradient steps	−1 (match train frequency)
Learning starts	100
Exploration (action) noise ξ	None

Table D.7

Hyperparameters of the client optimizers. These are held constant throughout all experiments.

	β_1	β_2	ϵ	λ	μ	τ
Adam	0.9	0.999	10^{-8}	0	—	—
SGD	—	—	—	0	0	0
SGDM	—	—	—	0	0.9	0

For the client optimizers, we only vary the learning rate, and use the default values for other hyperparameters. The default values are presented in Table D.7. For SGDM we set the momentum $\mu = 0.9$.

Appendix E. Sensitivity analysis

In this section, we perform a sensitivity analysis of the hyperparameters of both sets of experiments. We present the analysis of the client optimizers in Section E.1, and analyse the server optimizers in Section E.2.

E.1. Client optimizers

In our experiments comparing different client optimizers, we had two tunable hyperparameters: the local updates per round U and client learning rate η_l . First, we look at how the choice of U affects the performance of the federated agent. We present the progression of energy consumption and comfort violation on the evaluation environment for different values of U in Figs. E.7, E.8 and E.9, for Adam, SGD and SGDM, respectively. The performance for different values of U tends to be quite comparable, for every tested client optimizer. We do not observe any one value of U that consistently outperforms the others, though $U = 4$ tends to fall short of the others in terms of energy consumption. We notice that $U = 4$ also has slightly worse stability than other values of U , both with respect to energy consumption and comfort violation. This is in line with the previous experiments reported in Hagström [13], where also a centralized agent trained on data pooled from all environments (i.e., having $U = 1$) was shown to underperform against the federated agents.

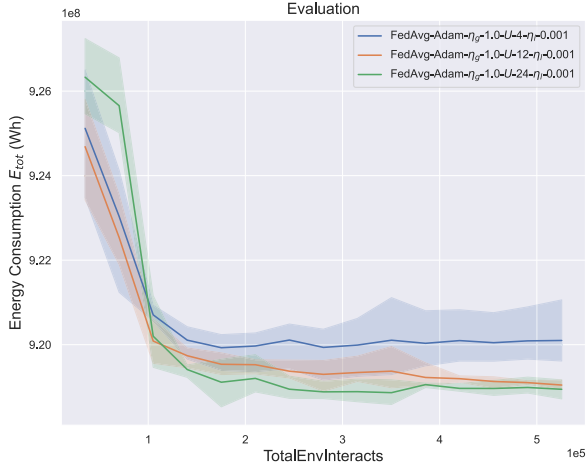
Considering these results, conclude that the federated agent is robust to the choice of U . It is, however, advisable to use larger values, not only because of the worse stability when performing global aggregation after every local update but also because larger values of U mean fewer communication rounds, reducing the communication costs of the federated algorithm.

Next, we focus on the client learning rate η_l . We present the progression of energy consumption and comfort violation on the evaluation environment for different values of η_l in Figs. E.10, E.11 and E.12, for Adam, SGD and SGDM, respectively. The performance of the federated agent is sensitive to the client learning rate. In Fig. E.10, we see that higher learning rates can significantly increase the energy consumption of the agent. It can also lead to complete failure in learning a comfortable policy, with $\eta_l = 0.1$ having 100 % comfort violation. The inverse relationship is true for SGD and SGDM, as can be seen in Figs. E.11 and E.12. They tend to achieve lower energy consumption with higher η_l and the choice of η_l seems to have less of an impact on the comfort violation.

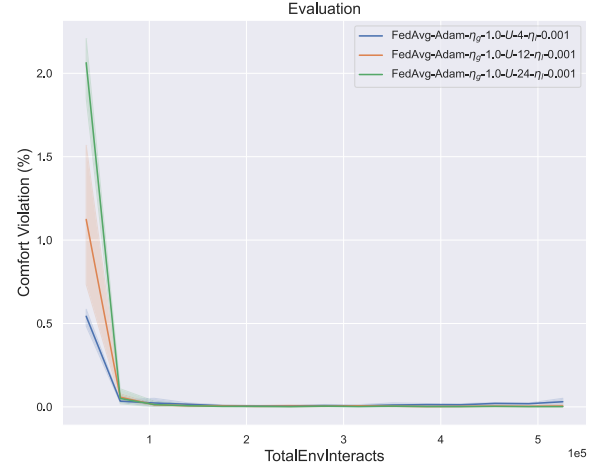
In Section 4.3.1, we concluded Adam to be the best choice of client optimizer. Based on the observed sensitivity to the client learning rate, it is advisable to use values of $\eta_l \leq 0.001$ for safe performance.

E.2. Server optimizers

We first consider the sensitivity of the FedAvgM algorithm to its hyperparameters. In Fig. E.13, we present the progression of energy consumption and comfort violation in the evaluation environment for different values of the global learning rate η_g . We notice a trend of improved performance for larger values of η_g , both in terms of energy consumption and comfort violation. The global learning rate also affects the learning

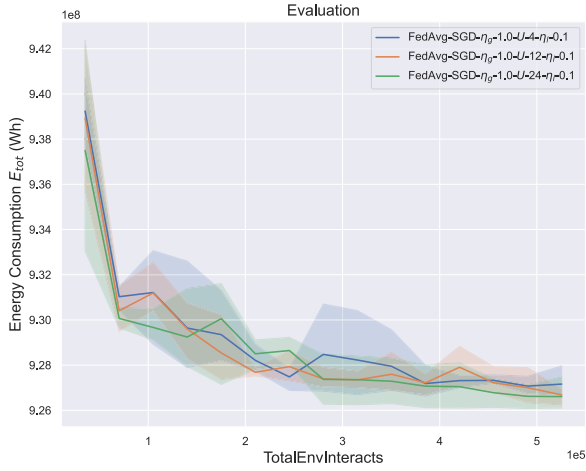


(a) Energy consumption

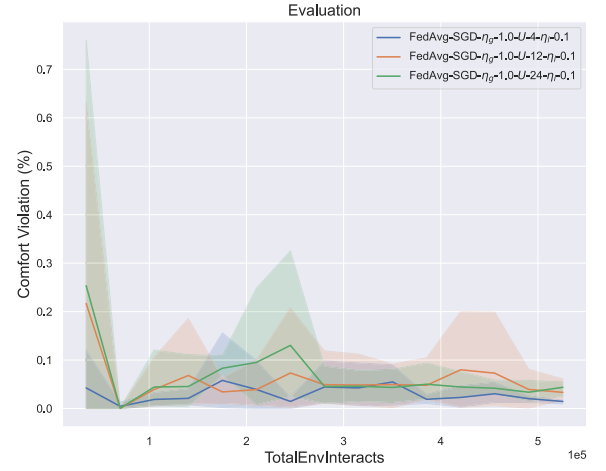


(b) Comfort violation

Fig. E.7. Comparing the performance of FedAvg with Adam as client optimizer on the evaluation environment for different local updates per round U . We fix $\eta_l = 0.001$.

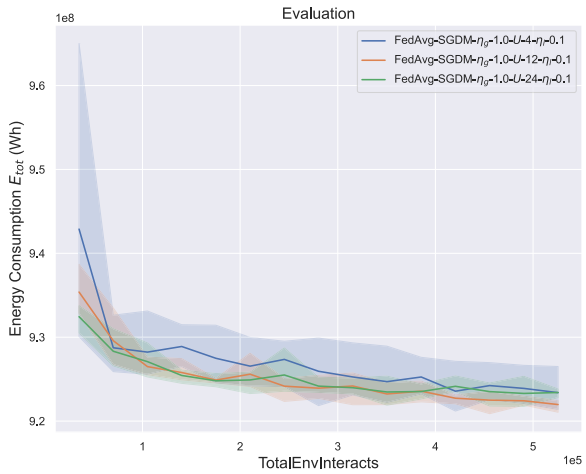


(a) Energy consumption

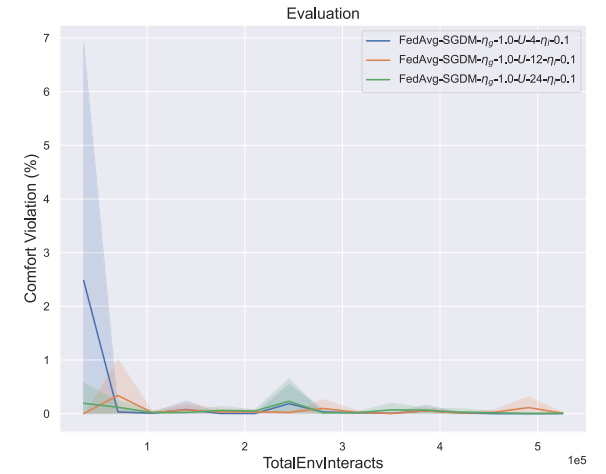


(b) Comfort violation

Fig. E.8. Comparing the performance of FedAvg with SGD as client optimizer on the evaluation environment for different local updates per round U . We fix $\eta_l = 0.1$.

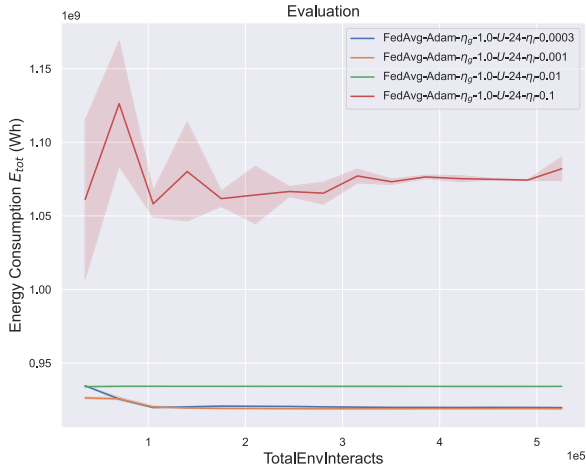


(a) Energy consumption

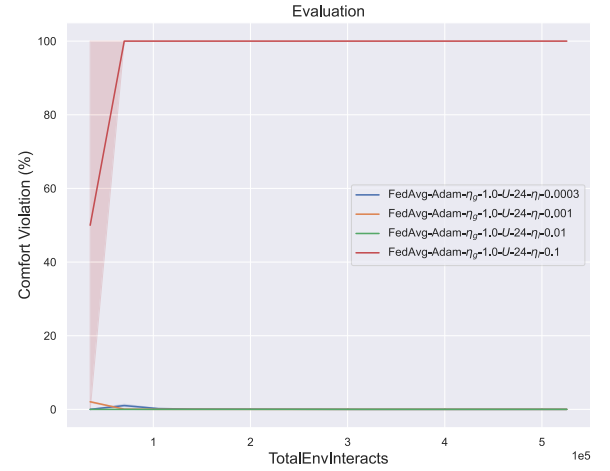


(b) Comfort violation

Fig. E.9. Comparing the performance of FedAvg with SGDM as client optimizer on the evaluation environment for different local updates per round U . We fix $\eta_l = 0.1$.

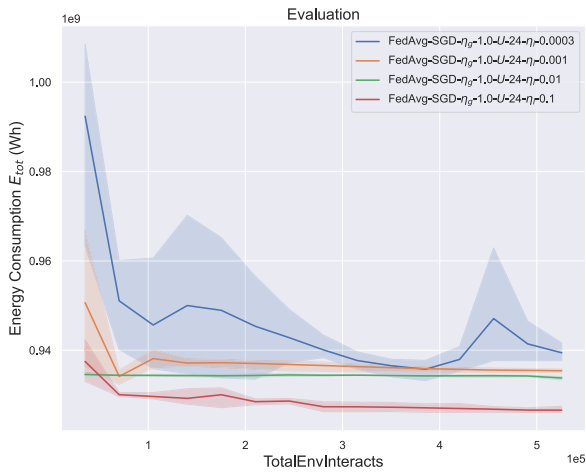


(a) Energy consumption

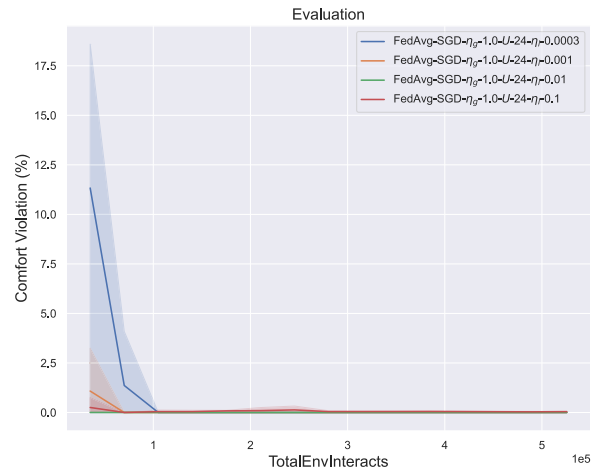


(b) Comfort violation

Fig. E.10. Comparing the performance of FedAvg with Adam as client optimizer on the evaluation environment for different client learning rates η_l . We fix $U = 24$.

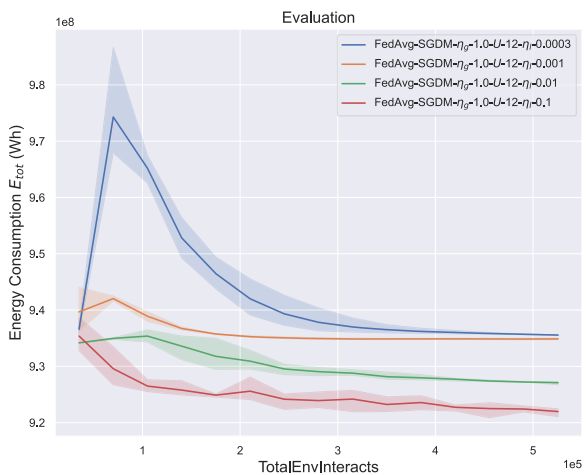


(a) Energy consumption

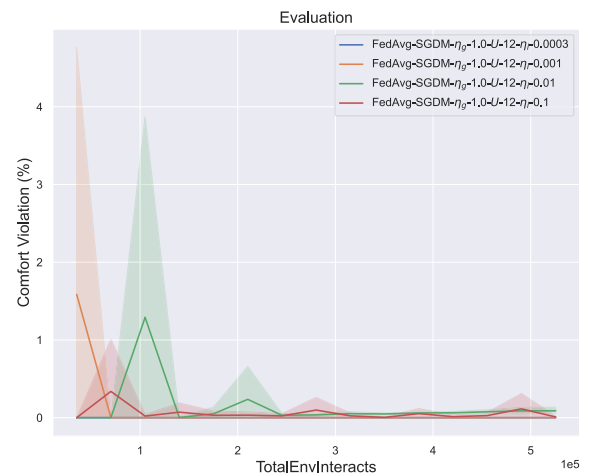


(b) Comfort violation

Fig. E.11. Comparing the performance of FedAvg with SGD as client optimizer on the evaluation environment for different client learning rates η_l . We fix $U = 24$.

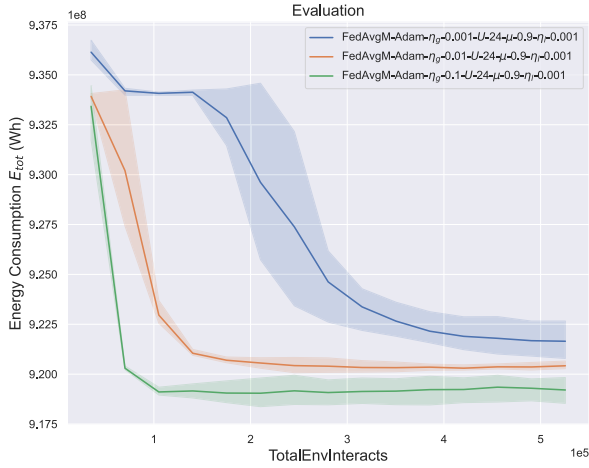


(a) Energy consumption

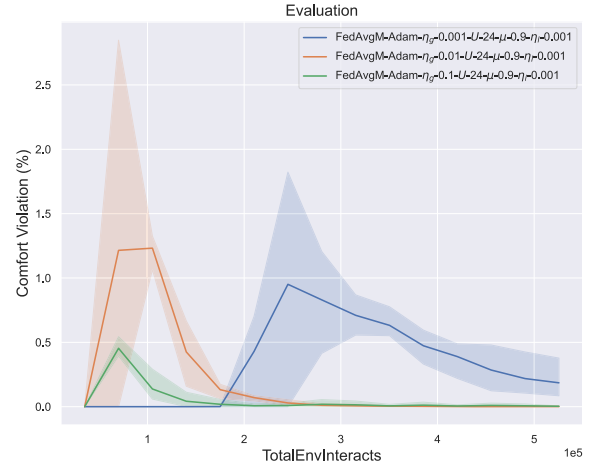


(b) Comfort violation

Fig. E.12. Comparing the performance of FedAvg with SGDM as client optimizer on the evaluation environment for different client learning rates η_l . We fix $U = 12$.

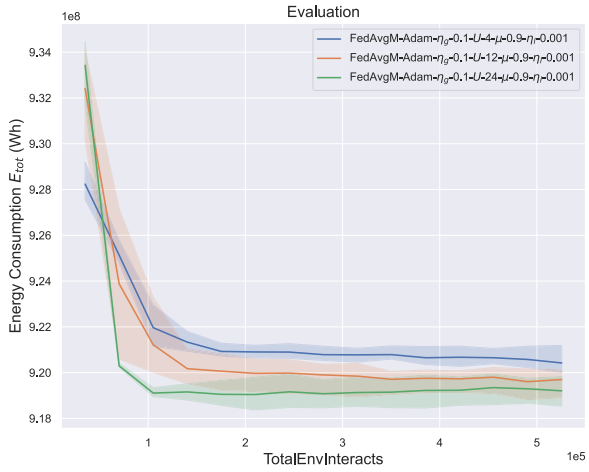


(a) Energy consumption

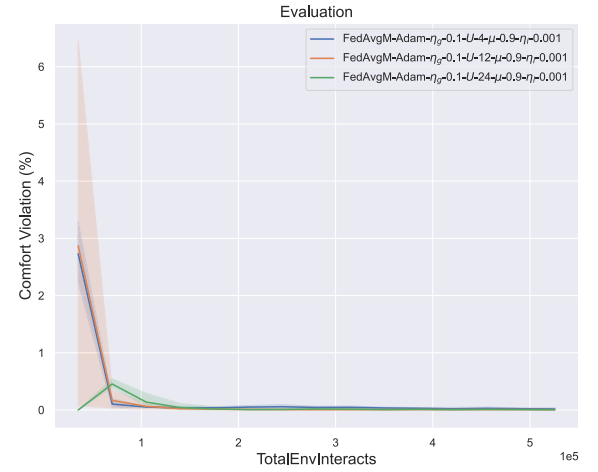


(b) Comfort violation

Fig. E.13. Comparing the performance of FedAvgM on the evaluation environment for different global learning rates η_g . We fix $U = 24$ and $\beta = 0.9$.

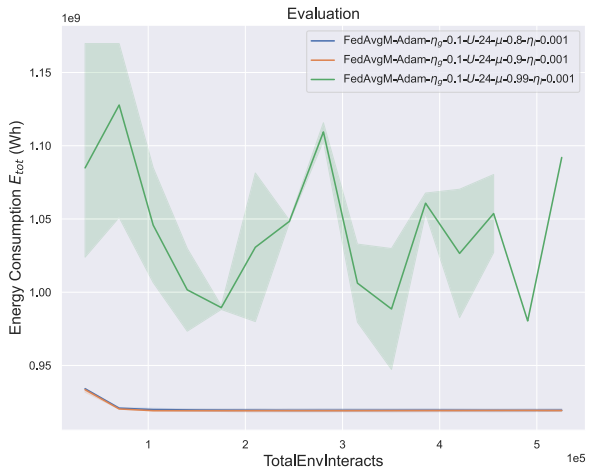


(a) Energy consumption

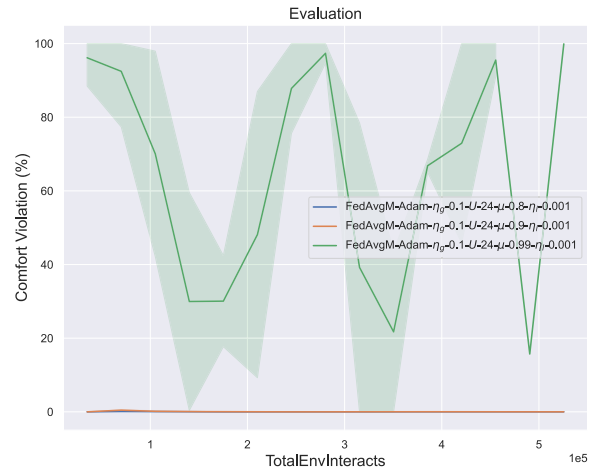


(b) Comfort violation

Fig. E.14. Comparing the performance of FedAvgM on the evaluation environment for different local updates per round U . We fix $\eta_g = 0.1$ and $\beta = 0.9$.

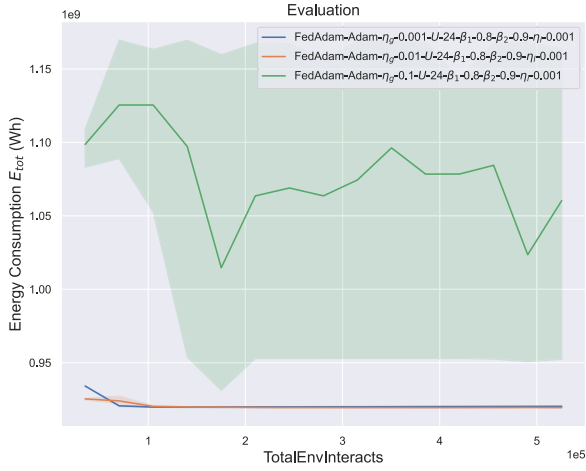


(a) Energy consumption

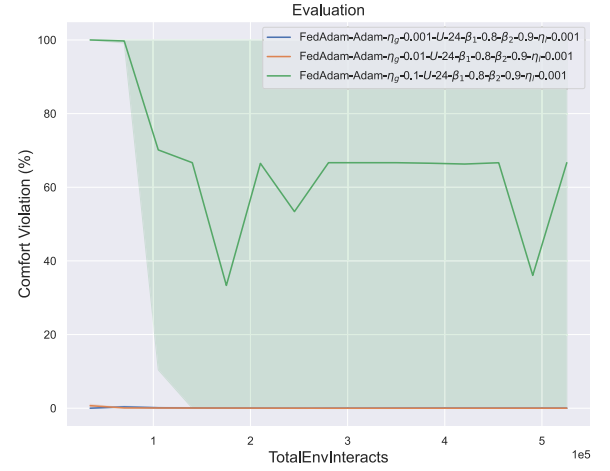


(b) Comfort violation

Fig. E.15. Comparing the performance of FedAvgM on the evaluation environment for different momentums β . We fix $\eta_g = 0.1$ and $U = 24$.

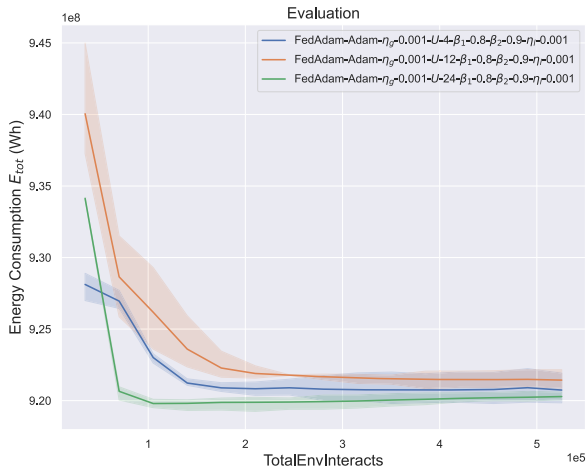


(a) Energy consumption

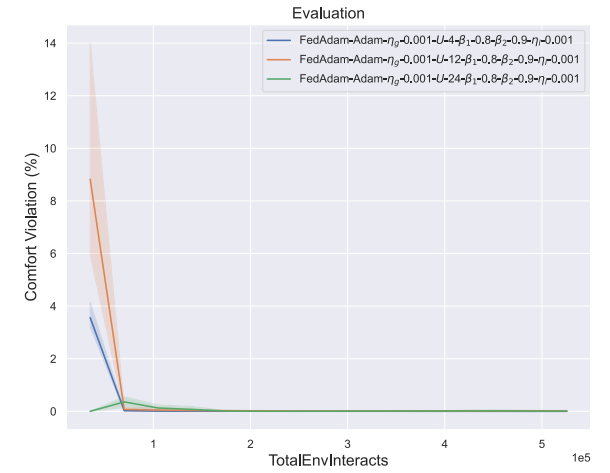


(b) Comfort violation

Fig. E.16. Comparing the performance of FedAdam on the evaluation environment for different global learning rates η_g . We fix $U = 24$, $\beta_1 = 0.8$ and $\beta_2 = 0.9$.

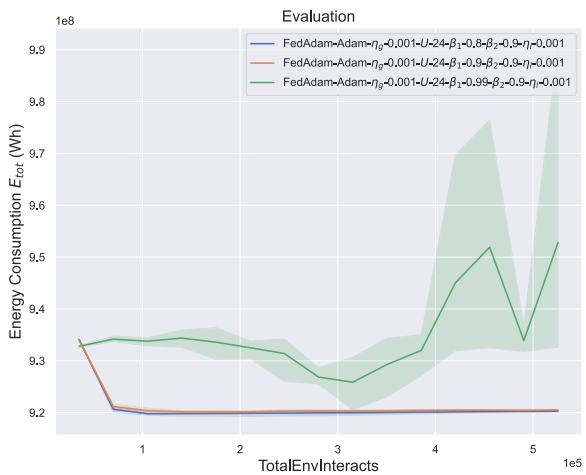


(a) Energy consumption

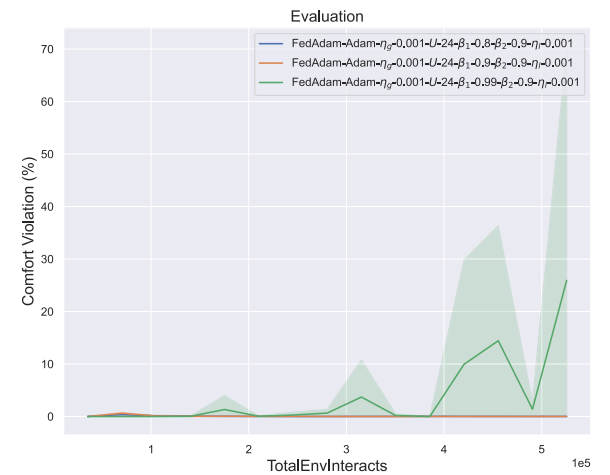


(b) Comfort violation

Fig. E.17. Comparing the performance of FedAdam on the evaluation environment for different local updates per round U . We fix $\eta_g = 0.001$, $\beta_1 = 0.8$ and $\beta_2 = 0.9$.



(a) Energy consumption



(b) Comfort violation

Fig. E.18. Comparing the performance of FedAdam on the evaluation environment for different β_1 . We fix $\eta_g = 0.001$, $U = 24$ and $\beta_2 = 0.9$.

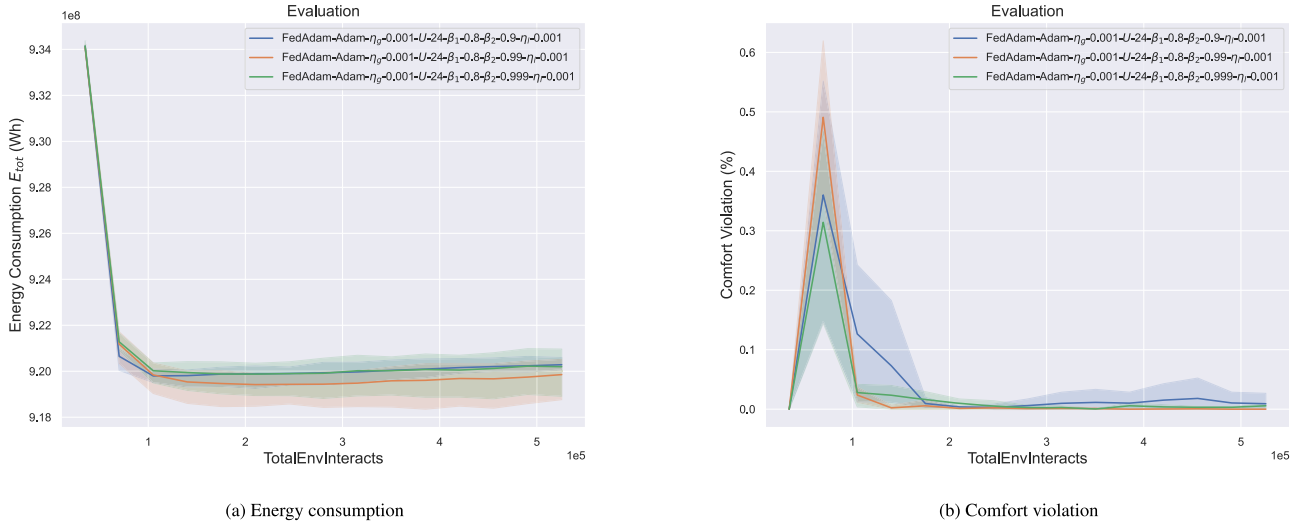


Fig. E.19. Comparing the performance of FedAdam on the evaluation environment for different β_2 . We fix $\eta_g = 0.001$, $U = 24$ and $\beta_1 = 0.8$.

speed and, to some extent, the learning stability. While increasing the learning rate tends to improve the learning speed and, thus, the performance of FedAvgM, one cannot use arbitrarily large values. We observed in our experiments that setting $\eta_g = 1.0$ tends to lead to exploding gradients, thus leading to an unusable policy. The FedAvgM algorithm is sensitive to the choice of global learning rate, and it needs to be chosen carefully for optimal performance. From Fig. E.14, we see that FedAvgM is less sensitive to the choice of U . FedAvgM displays similar performance, learning speed and learning stability for different values of U , though larger values perform slightly better in terms of energy consumption.

FedAvgM introduces the server momentum parameter μ . The progression of energy consumption and comfort violation for different values of μ are presented in Fig. E.15. The choice of μ has a considerable effect on the learning of the model. Too large a value leads to a significant increase in both energy consumption and comfort violation. The learning never converges, and the learning stability is significantly worsened, showing that the FedAvgM is also sensitive to the choice of momentum.

Similarly to FedAvgM, FedAdam is sensitive to the choice of global learning rate η_g , as can be seen in Fig. E.16. FedAdam, however, performs better with smaller learning rates. Larger learning rates lead to

a significant reduction in performance and learning stability, both in terms of energy consumption and comfort violation. Too large a global learning rate can also lead to failure to learn, as we observed that setting $\eta_g = 1.0$ to result in exploding gradients during training.

In Fig. E.17, we present the learning curves for different values of U . As with both FedAvg and FedAvgM, the performance, learning speed and stability are comparable for all tested values of U , and larger values display slightly improved energy consumption.

FedAdam has two adjustable moment parameters β_1 and β_2 . FedAdam seems to be more sensitive to the choice of β_1 than the choice of β_2 . In Fig. E.18, we see that too large a value of β_1 leads to a significant degradation in performance and learning stability. β_2 seems significantly more robust, with all tested values having comparable performance, learning speed and stability in terms of both energy consumption and comfort violation, as can be seen in Fig. E.19.

Appendix F. Additional plots

In Figs. F.20 and F.21 we show the training energy consumption and comfort violation curves for the environments omitted in Section 4.3.2.

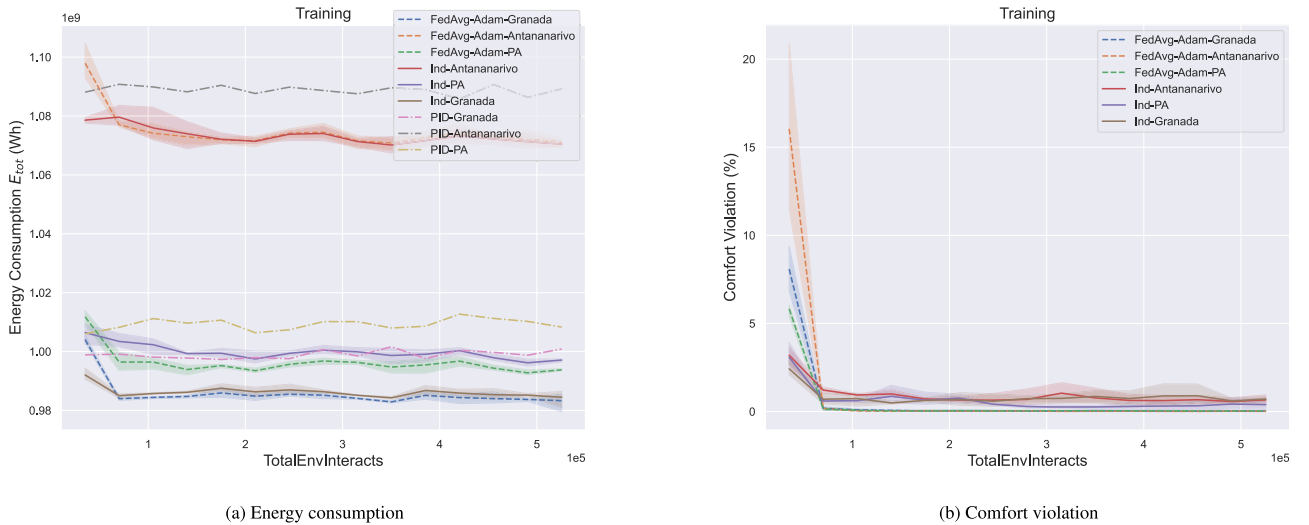


Fig. F.20. Progression of the energy consumption and comfort violation of FedAvg and independent agents on training environments Granada, Antananarivo and PA.

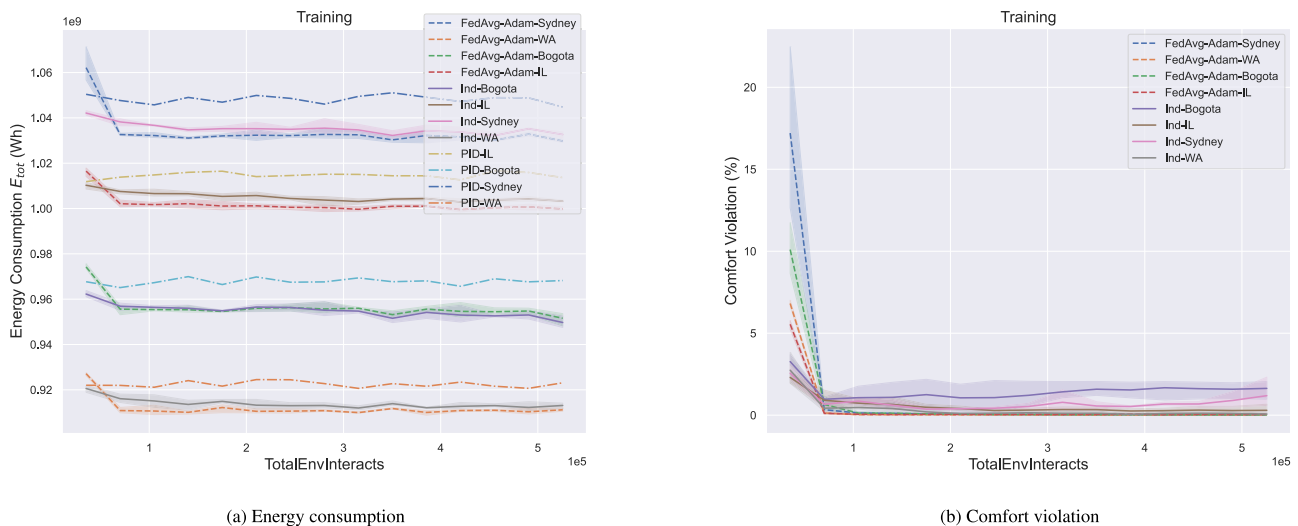


Fig. F.21. Progression of the energy consumption and comfort violation of FedAvg and independent agents on training environments Sydney, Bogota, WA and IL.

References

- [1] S. Fawzy, A.I. Osman, J. Doran, D.W. Rooney, Strategies for mitigation of climate change: a review, *Environ. Chem. Lett.* 18 (2020) 2069–2094.
- [2] M. Biemann, F. Scheller, X. Liu, L. Huang, Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control, *Appl. Energy* 298 (2021) 117164.
- [3] Z. Wang, T. Hong, Reinforcement learning for building controls: the opportunities and challenges, *Appl. Energy* 269 (2020) 115036.
- [4] A. Perera, P. Kamalaruban, Applications of reinforcement learning in energy systems, *Renew. Sustain. Energy Rev.* 137 (2021) 110618.
- [5] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 1273–1282.
- [6] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, (2014). *arXiv preprint arXiv:1412.6980*
- [7] T.-M.H. Hsu, H. Qi, M. Brown, Measuring the effects of non-identical data distribution for federated visual classification, (2019). *arXiv preprint arXiv:1909.06335*
- [8] S. Reddi, Z. Charles, M. Zahere, Z. Garrett, K. Rush, J. Konečný, S. Kumar, H.B. McMahan, Adaptive federated optimization, (2020). *arXiv preprint arXiv:2003.00295*
- [9] I. Tenison, S.A. Sreeramadas, V. Mugunthan, E. Oyallon, E. Belilovsky, I. Rish, Gradient masked averaging for federated learning, (2022). *arXiv preprint arXiv:2201.11986*
- [10] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: *International Conference on Machine Learning*, PMLR, 2018, pp. 1861–1870.
- [11] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, et al., Soft actor-critic algorithms and applications, (2018). *arXiv preprint arXiv:1812.05905*
- [12] T. Haarnoja, S. Ha, A. Zhou, J. Tan, G. Tucker, S. Levine, Learning to walk via deep reinforcement learning, (2018). *arXiv preprint arXiv:1812.11103*
- [13] F. Hagström, Using federated learning techniques to train deep reinforcement learning agents for HVAC control (2023).
- [14] J.R. Vázquez-Canteli, Z. Nagy, Reinforcement learning for demand response: a review of algorithms and modeling techniques, *Appl. Energy* 235 (2019) 1072–1089.
- [15] D. Weinberg, Q. Wang, T.O. Timoudas, C. Fischione, A review of reinforcement learning for controlling building energy systems from a computer science perspective, *Sustain. Cities Soc.* 89 (2023) 104351, <https://doi.org/10.1016/j.scs.2022.104351>.
- [16] C.W. Anderson, D.C. Hittle, A.D. Katz, R.M. Kretchmar, Synthesis of reinforcement learning, neural networks and PI control applied to a simulated heating coil, *Artif. Intell. Eng.* 11 (4) (1997) 421–429.
- [17] M.C. Mozer, The neural network house: An environment that adapts to its inhabitants, in: *Proc. AAAI Spring Symp. Intelligent Environments*, 58, 1998.
- [18] G.P. Henze, J. Schoenmann, Evaluation of reinforcement learning control for thermal energy storage systems, *HVAC&R Res.* 9 (3) (2003) 259–275.
- [19] S. Liu, G.P. Henze, Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory, in: *International Solar Energy Conference*, 47373, 2005, pp. 301–311.
- [20] S. Liu, G.P. Henze, Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory. Part 1. Theoretical foundation, *Energy Build.* 38 (2) (2006) 142–147.
- [21] S. Liu, G.P. Henze, Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: part 2: results and analysis, *Energy Build.* 38 (2) (2006) 148–161.
- [22] B. Sun, P.B. Luh, Q.-S. Jia, B. Yan, Event-based optimization with non-stationary uncertainties to save energy costs of HVAC systems in buildings, in: *2013 IEEE International Conference on Automation Science and Engineering (CASE)*, IEEE, 2013, pp. 436–441.
- [23] E. Barrett, S. Linder, Autonomous HVAC control, a reinforcement learning approach, in: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7–11, 2015, Proceedings, Part III 15*, Springer, 2015, pp. 3–19.
- [24] B. Li, L. Xia, A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings, in: *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, IEEE, 2015, pp. 444–449.
- [25] F. Ruelens, B.J. Claessens, S. Vandaal, B. De Schutter, R. Babuška, R. Belmans, Residential demand response of thermostatically controlled loads using batch reinforcement learning, *IEEE Trans. Smart Grid* 8 (5) (2016) 2149–2159.
- [26] G.T. Costanzo, S. Iacovella, F. Ruelens, T. Leurs, B.J. Claessens, Experimental analysis of data-driven control for a building heating system, *Sustain. Energy, Grids Netw.* 6 (2016) 81–90.
- [27] T. Wei, Y. Wang, Q. Zhu, Deep reinforcement learning for building HVAC control, in: *Proceedings of the 54th Annual Design Automation Conference 2017*, 2017, pp. 1–6.
- [28] M.A. Wiering, M. Van Otterlo, Reinforcement learning, *Adaptation, Learn., Optim.* 12 (3) (2012) 729.
- [29] M.J. Kochenderfer, T.A. Wheeler, K.H. Wray, Algorithms for Decision Making, MIT Press, 2022.
- [30] G. Gao, J. Li, Y. Wen, Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning, (2019). *arXiv preprint arXiv:1901.04693*
- [31] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, F. Li, Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning, *Appl. Energy* 281 (2021) 116117.
- [32] B. Chen, Z. Cai, M. Bergés, Gnu-rl: a practical and scalable reinforcement learning solution for building HVAC control using a differentiable mpc policy, *Front. Built Environ.* 6 (2020) 562239.
- [33] C. Gao, D. Wang, Comparative study of model-based and model-free reinforcement learning control performance in HVAC systems, *J. Build. Eng.* 74 (2023) 106852.
- [34] A. Nagy, H. Kazmi, F. Cheaib, J. Driesen, Deep reinforcement learning for optimal control of space heating, (2018). *arXiv preprint arXiv:1805.03777*
- [35] M. Khalil, M. Essegir, L. Merghem-Boulahia, Federated learning for energy-efficient thermal comfort control service in smart buildings, in: *2021 IEEE Global Communications Conference (GLOBECOM)*, IEEE, 2021, pp. 01–06.
- [36] M. Khalil, M. Essegir, L. Merghem-Boulahia, A federated learning approach for thermal comfort management, *Adv. Eng. Inf.* 52 (2022) 101526.
- [37] Y. Guo, D. Wang, A. Vishwanath, C. Xu, Q. Li, Towards federated learning for HVAC analytics: A measurement study, in: *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, 2020, pp. 68–73.
- [38] J. Gao, W. Wang, Z. Liu, M.F. R.M. Billah, B. Campbell, Decentralized federated learning framework for the neighborhood: a case study on residential building load forecasting, in: *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 453–459.
- [39] Y. Lu, L. Cui, Y. Wang, J. Sun, L. Liu, Residential energy consumption forecasting based on federated reinforcement learning with data privacy protection, *CMES - Comput. Model. Eng. Sci.* 137 (1) (2023) 717–732, <https://doi.org/10.32604/cmcs.2023.027032>.
- [40] Z. Wang, P. Yu, H. Zhang, Privacy-preserving regulation capacity evaluation for HVAC systems in heterogeneous buildings based on federated learning and transfer learning, *IEEE Trans. Smart Grid* 14 (5) (2023) 3534–3549, <https://doi.org/10.1109/TSG.2022.3231592>.

- [41] S. Lee, L. Xie, D.-H. Choi, Privacy-preserving energy management of a shared energy storage system for smart buildings: a federated deep reinforcement learning approach, *Sensors* 21 (14) (2021) 4898.
- [42] K. Fujita, S. Fujimura, Y. Sun, H. Esaki, H. Ochiai, Federated reinforcement learning for the building facilities, in: 2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS), IEEE, 2022, pp. 1–6.
- [43] R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2018.
- [44] J. Jiménez-Raboso, A. Campoy-Nieves, A. Manjavacas-Lucas, J. Gómez-Romero, M. Molina-Solana, Sinergym: a building simulation and control framework for training reinforcement learning agents, in: Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, Association for Computing Machinery, New York, NY, USA, 2021, p. 319–323. <https://doi.org/10.1145/3486611.3488729>.
- [45] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba, OpenAI Gym, (2016). [arXiv preprint arXiv:1606.01540](https://arxiv.org/abs/1606.01540)
- [46] A. TC, et al., Data center power equipment thermal guidelines and best practices, ASHRAE TC 9.9, ASHRAE, USA (2016).
- [47] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dormann, Stable-Baselines3: reliable reinforcement learning implementations, *J. Mach. Learn. Res.* 22 (268) (2021) 1–8. <http://jmlr.org/papers/v22/20-1364.html>.
- [48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch (2017).