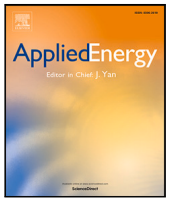
Applied Energy 347 (2023) 121364

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Performance, robustness, and portability of imitation-assisted reinforcement learning policies for shading and natural ventilation control

Bumsoo Park [a], Alexandra R. Rempel [b,*], Sandipan Mishra [a]

[a] Department of Mechanical, Aerospace, and Nuclear Engineering, Rensselaer Polytechnic Institute, Troy NY, 12180, United States of America
[b] Environmental Studies Program, University of Oregon, Eugene OR, 97403, United States of America

ARTICLE INFO

Keywords:
Imitation learning
Policy gradient reinforcement learning
Natural ventilation control
Shading control
Movable insulation control
Climate-responsive design

ABSTRACT

Space heating and cooling account for approximately half of all building-related energy consumption, emitting 3 Gt of CO_2 annually, or nearly 10% of the global total. Operable shading, natural ventilation, and solar heating are promising strategies for reducing these emissions, leveraging minimal mechanical energy to condition space with cool night air, cold night skies, and solar radiation. However, these strategies are under-utilized because their performance depends on rigorous coordination among their operable elements. Additionally, the individuality of such systems, and the lack of physics-based models suitable for control design, have thwarted the development of widely-applicable control strategies. To address this problem, here we develop a new data-driven strategy for the design of shading and natural ventilation controls in residential buildings using policy-based reinforcement learning (RL). To limit undesirable actions and reduce training time, we first used imitation learning to initialize RL training with expert knowledge, yielding an initial policy that reduced simulated late-spring space conditioning loads by ≥40% in 24 climatically diverse cities. This policy was then trained with RL in four cities representing Mediterranean, semi-arid, humid subtropical, and continental climates. When deployed in cities with unfamiliar yet related climates, these new policies reduced space conditioning loads by ≥50% in the humid subtropics and by ≥90% in the other three climates, showing exceptional portability. Further, their performance was unexpectedly robust to variations in dwelling orientation, glazing, internal heat gain, and air leakage. These results show the extraordinary potential of imitation-assisted RL in developing high-performance policies for dynamic passive heating and cooling control that remain effective in unfamiliar situations, removing a substantial barrier to passive systems advancement in carbon-free building operation.

1. Introduction

1.1. Significance of dynamic passive cooling and heating

Mechanical space conditioning is the single greatest energy end-use in buildings worldwide, consuming an estimated 25 PJ of energy and emitting approximately 3 Gt of CO_2 annually among International Energy Agency member countries alone [1], equaling nearly 10% of the global total [2]. As a result, the decarbonization of space heating and cooling is a high priority in efforts to avert the worst effects of climate change [3]. Over the past two decades, considerable advances have been made in building energy efficiency codes [e.g. 4], green building incentive programs [e.g. 5], urban landscape planning for cooling [e.g. 6], and renewable electricity supplies [7]. At the same time, increasing built area [8,9] and growing electricity demands [10, 11] are counteracting these advances, causing total space conditioning emissions to resist decline and showing that new approaches are urgently needed.

Dynamic passive cooling and heating have shown excellent potential to reduce cooling and heating loads in recent work [e.g. 12–17]. In these systems, operable elements such as shading devices, movable window insulation, vents, and window apertures allow buildings to capture or exclude, as desired, climatic resources such as solar heat and cool night air to reduce heating and cooling loads. For example, well-controlled natural ventilation and shading have shown the potential to reduce residential cooling loads by 50% or more in several Mexican climates [18]; by 70% or more across numerous Chinese cities [13]; and by up to 80% in the Pacific Northwest [15]. Similarly, passive solar heating with well-controlled movable insulation for windows has shown the potential to reduce residential heating loads by half or more

* Corresponding author.
E-mail address: arempel@uoregon.edu (A.R. Rempel).

https://doi.org/10.1016/j.apenergy.2023.121364
Received 29 November 2022; Received in revised form 10 May 2023; Accepted 29 May 2023
Available online 16 June 2023

Nomenclature

$\mathcal{D}$	Training dataset for imitation learning
π	Policy
π_θ	Approximated policy parameterized by θ
π_E	Expert policy
θ	Policy parameters
θ^*	Final policy parameters after reinforcement learning
θ_0	Policy parameters after imitation learning
a_t	Action of the MDP at time t
E_t	Sum of mechanical space heating and cooling loads at time t
G_t	Discounted cumulative rewards
r_t	Reward value of the MDP at time t
s_t	State of the MDP at time t
T_{in}	Indoor air temperature
T_{out}	Outdoor air temperature
ACH_{50}	Air changes per hour at 50 Pa
MDP	Markov Decision Process

in numerous U.S. climates [e.g. 16,17] and to contribute appreciably to space heating in southeastern Canada [14].

1.2. Passive system controls and limitations

In the majority of dynamic passive conditioning studies, the action of operable elements is signaled by static environmental setpoints such as indoor or outdoor air temperature values (detailed in Section 2.2.1). While such rule-based approaches are able to reduce cooling and heating loads in numerous climates, seasons, and building types [e.g. 17, 19–22], they have two important limitations. First, the setpoints used are typically optimized for particular spaces and climates [e.g. 19,20, 22,23]. As a result, recommended rules for the operation of shading, natural ventilation, and movable insulation vary widely, preventing establishment of the reliable, generally-applicable operational strategies needed to expand passive cooling and heating adoption. Second, once defined, rule-based systems cannot adapt to changes in weather or spatial configurations that may occur; for example, rules cannot adjust their setpoints to accommodate the fluctuations between overheating and overcooling that often occur during spring and fall months. These problems have highlighted the need for new control strategies that are responsive to weather variations, portable across climates, and robust to variations among the spaces they control, allowing them to perform well as soon as they are deployed. The purpose of this work is to develop a new method to create such controls.

1.3. Insights from advanced controls for mechanical systems

The development of advanced control strategies for space cooling and heating has focused primarily on mechanical, rather than passive, heating and cooling systems. Such work has taken both model-based approaches, relying on the derivation of dynamics models to solve for optimal control strategies, and model-free approaches, relying on the direct optimization of controllers without the formal use of models [24–26]. Model-based strategies such as model-predictive control (MPC) have become popular due to their excellent performance, particularly in buildings with sophisticated heating, ventilating, and air-conditioning (HVAC) systems [27]. However, deriving analytical models that are suitable for control design from fundamental physical laws is time-consuming and expensive, often prohibitively so at smaller residential scales [28]. This limits the appeal of model-driven methods for passive system applications.

Model-free methods such as reinforcement learning (RL) [29–31], in contrast, derive optimal control strategies through trial-and-error processes, without formal models of system dynamics, and are therefore less expensive to develop. Because of this trial-and-error nature, directly learning optimal control strategies from experiments on physical systems is not feasible. As a result, optimal RL strategies are typically learned in simulation environments, such as driving simulators for autonomous vehicles [32] and building energy simulations for building system controls [33]. The latter have shown excellent performance in controlling mechanical HVAC systems, reviewed by Sierla et al. [34]; among them, most have controlled high-level or end states, such as room air or supply air temperature setpoints, to reduce the number of control points [e.g. 31,33]. In passive heating and cooling systems, however, control actions must address the states of individual operable elements directly because of the situational relationships between element positions and their thermal influences. The retraction of window shading may promote solar heat gain (heating) during a winter day, for example, but heat loss (cooling) on a summer night. At the same time, these complex relationships cause passive systems to be excellent candidates for RL control, despite the necessary methodological departure from precedents established in mechanical HVAC systems [35–38].

1.4. State of the art and knowledge gap

The application of model-free RL to the development of control strategies for passive systems such as shading, natural ventilation, and direct solar heating is a newly emerging area, and published work is limited. Nevertheless, recent studies have shown that RL control of individual elements in passive systems can be highly successful, regularly out-performing rule-based systems, while highlighting the further challenges that remain [35,36,38,39]. The majority of these investigations have used tabular Q-learning or the related SARSA (state–action-reward-state–action) method for training RL agents. Cheng et al. [35], for example, used indoor illumination measurements and occupant preferences to train agents to control window blind positions and slat angles, saving cooling energy and improving visual comfort relative to manual and traditional automated controls. Similarly, Chen et al. [36] trained Q-learning RL agents to minimize thermal discomfort and space-conditioning energy use by assisting a mechanical HVAC system with natural ventilation, again finding that RL policies considerably improved performance compared to heuristic controls. In related work, Han et al. [39] trained RL agents using Q-learning and SARSA to control window opening to maintain thermal comfort and indoor air quality in a building without mechanical cooling or ventilation, easily out-performing manual controls. Likewise, our own investigations have shown Q-learning to be effective in developing control policies for integrated shading and natural ventilation, allowing these passive strategies to reduce cooling loads significantly in several climate types [38].

Further progress with tabular Q-learning methods will be constrained, however, by the inability of Q-learning to determine similarities among related states or actions, limiting its effectiveness in extensive state and action spaces such as those found in dynamic passive space-conditioning systems. Additionally, tabular Q-learning cannot address the continuous state and action spaces that appear in such systems. Although nonlinear function approximators such as neural networks can assess the proximities of states or actions by directly learning the Q-functions, as in deep Q networks (DQN) [40,41], the need to represent large numbers of states and actions explicitly induces lengthy training times in building control applications [37]. To address the problems posed by tabular methods, Ding et al. [37] developed a new neural architecture, the Branching Dueling Q-Network, for application to comprehensive building control, including shading and natural ventilation; these trained policies predicted July energy savings of about 18% in simulations of a non-residential building in two hot-summer climates. Policy-gradient RL algorithms such as

REINFORCE [42] also appear promising: our own preliminary studies have shown that REINFORCE-trained policies reduced early-summer space conditioning loads appreciably (≥45%), in residential building simulations in a range of climates, through the control of shading, movable insulation, and natural ventilation [38].

Each of the model-free studies above noted significant problems, as well: training RL agents to reach optimal performance required explorations of the design space that were time-consuming [36,38], resource-intensive [39], or created so much discomfort in occupied space as to be infeasible in practice [35]; several also noted the need to train agents for the specific building and climate in which deployment was to occur [37–39]. An important knowledge gap therefore remains regarding the design of RL policies for passive cooling and heating control that (i) learn rapidly, overcoming time and resource limitations; (ii) show minimal undesirable behavior during learning, avoiding actions that cause discomfort; (iii) provide excellent portability across climates; and (iv) maintain consistent performance among building variations, relaxing the specificity of trained agents for particular situations. Addressing these demanding features of RL, such that trained policies are ready for general deployment in occupied spaces, is the central goal of this work.

1.5. Approach and original contribution

To address the first challenge, that of creating RL agents that are able to learn rapidly and show minimal undesirable behavior during training, here we explore the potential of warm-starting the agents with existing expert knowledge governing shading and natural ventilation. The intent of this approach is to create an initial policy that is closer to the optimal policy than a randomly chosen guess, reducing excursions into undesirable actions as well as requirements for time and data during training. Next, to investigate the challenge of creating policies that are portable among climates, the initialized RL policy is trained in several distinct climate types, generating climate-adapted policies. These policies are then evaluated in climatically related yet unfamiliar locations, without further refinement or knowledge of system dynamics, to reveal their portability. Finally, addressing the challenge of creating policies that are robust to spatial variations, the climate-adapted policies are evaluated in dwellings that vary in orientation, window glazing, internal heat gain, and air leakage. Such robustness is beyond the capability of rule-based controls for shading and natural ventilation, and to our knowledge, it has not been examined among RL control policies for these systems.

This work provides three original contributions: first, it represents the first successful attempt to encode expert knowledge of passive space cooling and heating control into an initial RL policy through imitation learning. Second, it represents the most climatically extensive investigation of RL policies for shading and natural ventilation control to date; to our knowledge, it is also only the second study [38] of policy-gradient RL applied to such systems. Third, it provides the first evidence that RL policies trained in this way can reduce space conditioning loads extensively and consistently across climatic regions and across typical building variations, showing excellent potential to reduce space conditioning loads without climate- or space-specific training.

2. Problem formulation and preliminaries

2.1. Problem statement

The objective of this work is to develop a control strategy that uses acquirable environmental measurements to actuate shading, window apertures, and movable insulation in a building to reduce space cooling and heating loads, as illustrated in Fig. 1, while avoiding unproductive actions. To address the knowledge gap above, the control algorithm must have the following features: (i) the ability to learn from limited data, requiring minimal modeling and/or commissioning effort; (ii) the ability to use data from available sensors to yield appropriate decisions among broadly related climates, showing portability; and (iii) general applicability to building variations beyond those used in training, showing robustness. We leverage access to a high-fidelity simulator (EnergyPlus) to design this control strategy in a data-driven model-free manner. In this study, we assume that we can measure (observe) indoor and outdoor air temperatures, heat flux through window surfaces (as demonstrated in [43]), and space heating and cooling loads at each control update. In model-free data-driven control design, performance relies on the quality as well as the quantity of training data. Since the data available for training are often limited, whether derived from simulations or field measurements, the intelligent integration of expert knowledge into the design, detailed below (Section 2.2) is an attractive option for reducing the need for training data.

2.2. Development of the expert policy

2.2.1. Precedent control rules

To date, most passive cooling and heating control studies have employed rule-based strategies in which the operation of shading, windows, or movable insulation is triggered by an environmental condition. In window shading control, a central dilemma has been the choice of the most effective environmental parameter(s) for signaling operation. Because the majority of studies have addressed workplaces, which require visual comfort, illumination values and glare indices have become popular in meeting cooling goals as well [e.g. 20,23,44–47]. These setpoints have been chosen to minimize electric lighting loads [e.g. 44]; to maintain workplane illuminance within desired ranges [e.g. 44,45]; to minimize glare [e.g. 20,44,46]; and to minimize the sum of cooling, heating, and lighting energy consumption [e.g. 23, 46]. The primary disadvantage of illumination- and glare-based controls for cooling, however, is the lack of close correlation between illumination and solar heat gain [e.g. 43], giving effective setpoints high specificity to particular spaces, orientations, and climates [23,46].

Studies interested primarily in cooling have therefore often controlled shading according to incident solar radiation upon window surfaces and/or transmitted solar radiation through window glazing [e.g. 21]. These strategies have shown excellent cooling performance, but they have had difficulty switching between heating and cooling modes, which often alternate during spring and fall months [e.g. 19, 46]. In response, investigators have adopted separate shading setpoints tailored to the two modes [e.g. 46,48] or to specific indoor and/or outdoor air temperature ranges [21,48,49], confining shading actions to hours when cooling was desired. Although these approaches have improved performance, their space-specificity has persisted. To address this problem, recent studies have investigated window surface heat flux, a promising metric indicating net window heat gain or loss independent of space, orientation, or climate [e.g. 43,48]. Additionally, the emergence of new, low-cost heat flux sensors, combined with straightforward methods to filter out sensor noise [43], have allowed heat flux sensing to become an affordable and effective option for cooling-focused shading control. We therefore use window heat flux measurements to control shading in the expert policy.

The primary challenge in natural ventilation control, in contrast, has been the great sensitivity of indoor air temperature to outside air temperature when ventilation is active, causing spaces to overcool easily and inducing window opening–closing oscillations [e.g. 28, 50,51]. In response, investigators have studied numerous combinations of setpoints and deadbands involving indoor air and/or operative temperatures, outdoor air temperatures, and occupancy status, often varying by season [e.g. 15,19,21,50]. These efforts have been highly successful in both residential and non-residential buildings, diminishing cooling loads by substantial fractions. Additionally, the resulting rule systems have shown relative consistency across climates and seasons, as reflected in the expert policy.

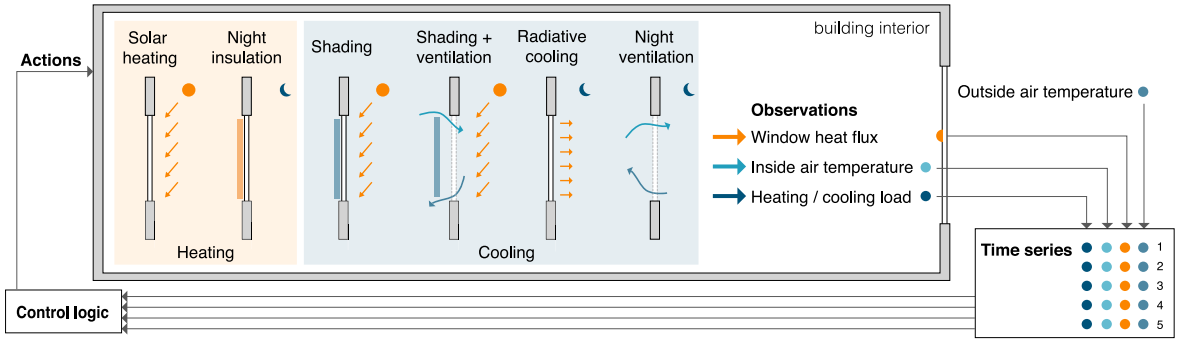


Fig. 1. Feedback control of operations in the dynamic passive system of interest. Observations included inside and outside air temperatures, window surface heat flux values, and cooling or heating loads at each 15-min timestep. The learning goal was to develop a control logic for operating shading, night insulation, and window apertures to minimize the sum of space cooling and heating loads.

Efforts to control shading and natural ventilation simultaneously have met compounded challenges, however, because the effects of shading and natural ventilation interact: for example, shading reduces heat storage in wall and floor materials, causing a shaded space to cool more rapidly than an unshaded one during natural ventilation. These challenges have grown when passive cooling systems have been expanded to include movable insulation for passive heating, leading researchers to propose highly involved rule systems [21,52] or to seek effective control setpoints through parametric searches [19] or numerical optimizations [17]. Although the resulting strategies have remained space-specific, they have confirmed the extent of passive heating and cooling potential in these spaces, often meeting 40%–80% of the respective loads, and they further informed the establishment of setpoints for the consensus expert policy (Section 2.2.2).

2.2.2. Expert policy

In the expert policy π_E, window heat flux was chosen as the most universally effective environmental parameter for shading control, for cooling, from the evidence of Danis et al. [43]. These cooling-focused rules were modified, however, to retract shading (i.e., to allow solar heat gain to occur) when indoor air was sufficiently cool. A deadband was also provided to limit oscillation. Setpoints defining these modifications were found heuristically, by automated trial-and-error, to minimize space-conditioning loads in cities representing four distinct climate types (Section 3.1) while limiting shade oscillation. In the final policy π_E (Table 1), shading was extended if the indoor air temperature T_{in} equaled or exceeded a threshold of 18.5 °C in any of the 5 previous 15-minute timesteps, representing an elapsed hour, and the window gained heat in the current timestep; retracted if the indoor air equaled or exceeded 18.5 °C at any time in the past hour and the window lost heat definitively (<-2 W/m^2) in all 5 previous timesteps; and maintained its current status otherwise.

Natural ventilation control, in turn, followed the summer rule structure of Schulze and Eicker [50]: ventilation was enabled only when indoor air was sufficiently and consistently warm and when outdoor air was sufficiently cool; once enabled, the window aperture opening, as a proxy for air exchange, increased with outdoor air temperature until the outdoor air became too warm. Specifically, natural ventilation was enabled only if T_{in} equaled or exceeded the threshold of 18.5 °C in *all* of the 5 previous timesteps and if the outside air temperature T_{out} measured 25 °C or cooler in the current timestep; once enabled, the extent of aperture opening was modulated according to T_{out} to minimize overcooling. As above, setpoints defining these boundaries were found heuristically to minimize late-spring space-conditioning loads in the four representative cities. Because space conditioning needs were initially dominated by cooling, explicit movable insulation control for heat retention was not included. However, the shading element possessed insulating properties, allowing it to serve as movable insulation (see Section 3.1 and Table 2), and RL policies learned to use it in this capacity (see Section 4.2). The use of this expert policy to initialize RL training is described in Section 2.4.

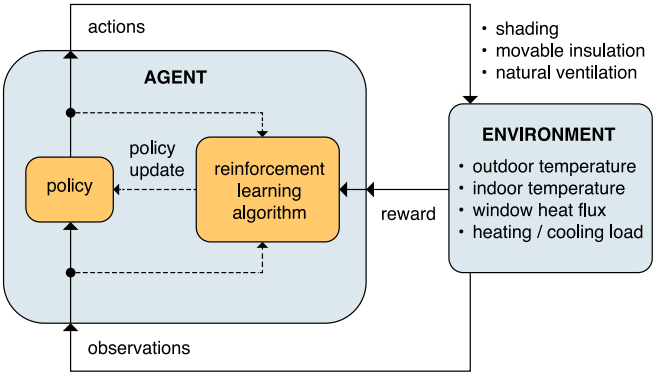


Fig. 2. Reinforcement learning process. The purpose of reinforcement learning is to train an agent to perform desired actions in an unfamiliar environment through a series of interactions with that environment. In each cycle, the agent receives observations from the environment; responds with actions that influence the environment; and receives a reward that indicates the relative success of the actions. The learning algorithm then updates the policy parameters based on the observations, actions, and reward, with the goal of maximizing the cumulative reward received.

2.3. Reinforcement learning

Reinforcement learning (RL) refers to a class of machine learning algorithms in which a classifier or regressor is trained to make a sequence of decisions to maximize a cumulative reward. The learner in RL is typically termed the *agent*, which learns an appropriate control strategy (or policy) by interacting with its surroundings, known as the *environment*, as illustrated in Fig. 2. The agent uses the control strategy to decide on an action, while the environment provides observations (or states) and returns rewards to the agent. During the training process, the agent takes an action, observes the outcome, and is given a reward as a consequence of the action and the outcome. Based on the action, the observation, and the reward, the agent updates its current control strategy to maximize expected cumulative future rewards. This process is repeated until the agent learns a strategy to achieve the goal (interpreted as maximizing the reward) in a dynamical environment.

2.3.1. Markov decision processes and reinforcement learning problem formulation

The design of RL algorithms is built on modeling the underlying system as a Markov Decision Process (MDP). An MDP consists of the following elements: the possible states S of the system, the allowable actions $\mathcal{A}$, the rewards $R_t \in \mathcal{R}$, and the transition probabilities $P(\cdot|\cdot,\cdot)$ between states, given an action. The states (S_t) are formal representations of the observations (measurements) from the environment at

Table 1
Expert policy.

	Condition 1	Duration[a]	Condition 2	Duration	Action[b]
Shading	$T_{in} \geq 18.5$ °C	any	WHG[c] > 0 W/m^2	any	1 (fully on)
			WHG < -2 W/m^2	all	0 (fully off)
			-2 W/m$^2 \leq$ WHG ≤ 0 W/m^2	all	Keep previous action
	$T_{in} < 18.5$ °C	any	None	n/a	Keep previous action
Natural ventilation	$T_{in} \geq 18.5$ °C	all	$T_{out} \leq 11$	current	0.01
			$11 < T_{out} \leq 14$		0.01 T_{out}
			$14 < T_{out} \leq 16$		0.02 T_{out}
			$16 < T_{out} < 18.5$		0.05 T_{out}
			$18.5 \leq T_{out} \leq 25$ and $T_{out} < T_{in}$		1 (fully open)
			$25 < T_{out}$		0 (fully closed)
	$T_{in} < 18.5$ °C	any	None	n/a	0 (fully closed)

[a]Number of timesteps in the past hour.
[b]Shading fraction on or window fraction open.
[c]Window heat gain.

time t; the action A_t describes the effect of the agent on the environment; and the reward R_t is a scalar value returned to the agent when progressing from S_t to S_{t+1}, taking action A_t. The transition probability $P(S_{t+1}|S_t, A_t)$ is the probability of transitioning from S_t to S_{t+1} by taking the action A_t. The transition probabilities capture the inherent dynamics of the underlying environment; here, these include the dwelling and its operable elements.

Given an MDP, the policy (of the agent) is a function that maps the states $\mathcal{S}$ to actions $\mathcal{A}$. Policies can be deterministic ($\pi : \mathcal{S} \rightarrow \mathcal{A}$) or stochastic ($\pi : \mathcal{S} \rightarrow \Pr(\mathcal{A})$) in nature. The policy is designed to maximize a cumulative reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{\,t} R(S_{t+1}, A_t)\right]. \tag{1}$$

This optimal control problem, though computationally intensive, can be directly solved if the transition probabilities are perfectly known. In a model-free learning scheme such as RL, however, the transition probabilities of the MDP are *unknown*; as a result, the optimal policy is found through interaction with the environment [29].

2.3.2. Policy-based reinforcement learning

The goal of an RL algorithm is to determine the optimal policy π^* that maximizes the expected cumulative *future* reward, $\mathbb{E}_\pi\left[G_t\right]$, where $G_t = \sum_{k=t+1}^{\tau} \gamma^k r_k$. Here, $\gamma \in [0, 1]$ is a discount factor on future rewards; G_t is the discounted cumulative future reward; and $\mathbb{E}_\pi\left[G_t\right]$ is the expectation of G_t when actions are determined from the policy π. Two broad classes of RL algorithms exist for determining the optimal policy π^*, namely, value-based RL and policy-based RL methods [29].

For a given policy π, the *value function* is the expected cumulative future reward, given a state s at time t, $V^\pi(s)$:

$$V^\pi(s) \equiv \mathbb{E}_\pi\left[G_t \mid S_t = s\right] \tag{2}$$

In policy-based RL, the optimal policy is directly learned without determining the value function explicitly. The policy is functionally represented by a set of d basis functions $\mathcal{B}$ and an associated parameter vector $\theta \in \mathbb{R}^d$, i.e., $\pi \equiv \pi_\theta$. The parameter θ is updated through gradient ascent in policy-gradient (PG) methods, with the objective of maximizing the expected future rewards. Hence the cost function $J(\theta)$ for an episodic case is defined as the state-value function, $J(\theta) = V^{\pi_\theta}(s_0)$, where s_0 is the initial state at the beginning of the episode; i.e.,

$$\theta^* = \arg\max_{\theta \in \mathbb{R}^d} J(\theta) = \arg\max_{\theta \in \mathbb{R}^d} \mathbb{E}\left[V^{\pi_\theta}(s_0)\right], \tag{3}$$

where an episode refers to the length of a single 'run', e.g., the time horizon over which the task is being executed. If the gradient of the cost with respect to the policy parameters is available, gradient ascent can be used to find a locally optimal policy. According to the policy gradient theorem [29], the gradient of the cost $\nabla_\theta J(\theta)$ can be written as

$$\nabla_\theta J(\theta) = \mathbb{E}\left[G_t \nabla \ln \pi(A_t \mid S_t, \theta)\right]. \tag{4}$$

To determine this gradient, here we employ the well-known REINFORCE [42] algorithm, a policy-gradient approach that uses a Monte-Carlo estimator for the gradient, i.e., an empirical *estimate* of the expectation of the gradient is used to compute the parameter update. The expectation term in Eq. (4) is computed from empirical reward samples from trajectories consisting of state, action, and reward value triplets. Algorithm 1 outlines the general REINFORCE algorithm.

Algorithm 1 REINFORCE: Monte-Carlo PG

Initialize parameter vector $\theta \in \mathbb{R}^d$
for episode = 1,…,N **do**
  Generate trajectory $s_0, a_0, r_1, \ldots, s_{\tau-1}, a_{\tau-1}, r_\tau$, following π_θ
  for each timestep of the episode, $t = 0, \ldots, \tau - 1$ **do**
    $G_t \leftarrow \sum_{k=t+1}^{\tau} \gamma^{k-t-1} r_k$
    $\theta \leftarrow \theta + \alpha \gamma^t G_t \nabla_\theta \ln \pi(a_t \mid s_t, \theta)$
    until $\gamma^t G_t \nabla_\theta \ln \pi(a_t \mid s_t, \theta)$ is small enough
  end for
end for

2.4. Imitation learning

Encoding expert knowledge into an approximated policy (π_E) that serves as the initial guess for an RL agent can reduce training time and data requirements. One approach to embed this expert knowledge is through imitation learning, in which a classifier is used to learn the behavior of an expert to perform complex tasks [53–55]. This is accomplished by training the classifier to predict, or classify, the action of the expert given an observation (measurement). These classifiers are trained using supervised learning, such that the prediction error is minimized for a collection of given observation-action pairs.

Relying on pure imitation learning has some limitations, however. One problem is that the assumption of independent identical distribution (i.i.d.) among the states in the training dataset may not always be satisfied, leading to unreliable performance and poor learning outcomes [56]. This occurs because the state distributions encountered by the learner and the expert are different: not all decisions of the learner are identical to those of the expert, causing the state trajectories to deviate from those experienced by the expert. Typically, this is addressed through iterative training approaches, in which the expert provides feedback during the training process to help the apprentice learn to recover from mistakes and to extrapolate the expert's behavior in unforeseen scenarios [e.g. 57–59]. Additionally, the expert policy may perform poorly in some scenarios; if so, relying on imitation learning alone is likely to yield similarly poor performance for those scenarios. We address these issues below (Section 3.3.1).

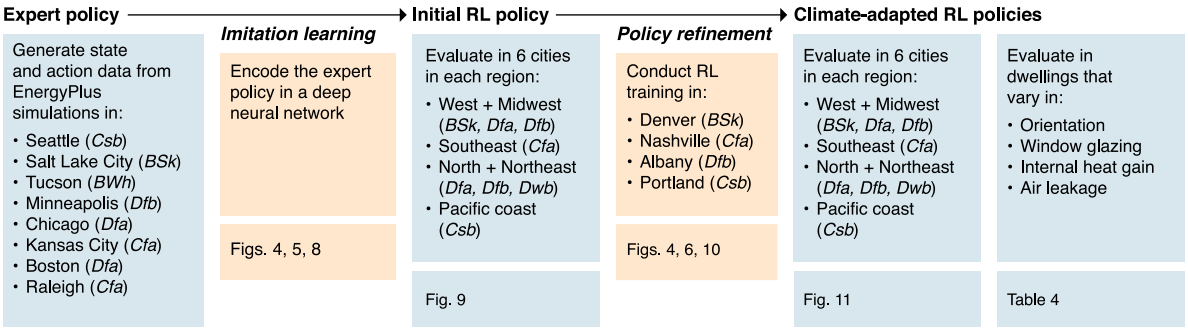


Fig. 3. Workflow. An expert policy was used to generate state–action data to support imitation learning, yielding an initial RL policy with abilities comparable to those of the expert. Refinement then occurred through RL training in multiple cities, yielding climate-adapted policies for further evaluation of performance among unfamiliar climates and building variations.

3. Methods

The methods described below followed the sequence of steps shown in Fig. 3. First, a dwelling model with operable windows and shading (Section 3.1) was simulated under the control of an expert policy (Section 2.2.2), in eight climatically diverse cities, to generate paired state–action data. Next, guided by these data, an imitation learning algorithm was used to mimic the expert policy and to generate the initial RL policy (Section 3.3.1). The initial RL policy was then evaluated by dwelling simulations in numerous additional cities, grouped into four broad climatic regions. Next, the initial RL policy was refined through training in cities representing each of the four climatic regions to create four climate-adapted policies (Section 3.3.2). Climate-adapted RL policies were then evaluated for portability within their corresponding climatic regions and for robustness among dwelling variations.

3.1. Building energy modeling

Passive cooling and heating control policies were developed using simulations of a South-facing one-bedroom apartment, represented as a single thermal zone in EnergyPlus v9.2 [60]. Model geometry was specified in Euclid 0.9.4.2 [61], an extension for SketchUp (Trimble Inc., 2017), and glazing properties were specified in WINDOW 7.7 [62] and exported as spectral IDF objects for use in EnergyPlus. For simplicity, opaque envelope and glazing materials, infiltration, and internal heat gain rates were specified to comply with requirements for a single climate, Climate Zone 5, of the 2018 International Energy Conservation Code [4] (Table 2); however, typical variations in these factors were considered as well (Section 4.5). Space heating and cooling were provided by `IdealLoadsAirSystem` objects, yielding estimates of heating and cooling loads independent of equipment efficiency; heating and cooling thermostat setpoints were set at 18 °C and 25 °C, respectively, again consistent with 2018 International Energy Efficiency Code guidelines [4]. All simulations used 15-min timesteps.

Ventilation for fresh air was maintained at 0.007 m^3/s, consistent with the recommendations of ASHRAE Standard 62.2-2019 [63], and natural ventilation airflow was calculated independently of this quantity and added to it. Window area available for natural ventilation totaled 5.1 m^2, distributed between two windows. Natural ventilation was simulated with `ZoneVentilation:WindAndStackOpenArea` objects to improve processing speed without compromising reliability in this simple case [15]. Shading and movable insulation, in turn, were specified with `WindowShadingControl` objects, in which shading and movable insulation were represented by the same interior operable panels due to the inability of EnergyPlus 9.2 to support the action of multiple devices upon a single window [64].

Table 2
Building parameters.

Element	Properties
Dwelling	
Floor area	44.9 m^2 (483 ft^2)
Exterior façade area	20.9 m^2 (221 ft^2)
Window to wall ratio	24.7% (building); 7% (dwelling)
Exterior façade orientation	South
Location	Top floor
Glazing	
Area	5.1 m^2 (54.5 ft^2)
Properties	U=1.61 W/m^2K (0.28 Btu/h ft^2 °F); SHGC=0.55
Shading	
Optical properties	T_{vis} = 0.1; R_{vis} = 0.8
Thermal properties	U = 2.8 W/m^2K (0.5 Btu/h ft^2 °F, or 'R-2')
Edge conditions	Side tracks (opening multipliers = 0)
Exterior wall assembly	U = 0.32 W/m^2K (0.057 Btu/h ft^2 °F)
Roof assembly	U = 0.15 W/m^2K (0.026 Btu/h ft^2 °F)
Interior wall assembly	U = adiabatic; heat capacity = 390 kJ/m^3K
Floor assembly	U = adiabatic; heat capacity = 2010 kJ/m^3K
Internal heat gain	8.6 W/m^2 (0.8 W/ft^2)
Air exchange	
Ventilation for fresh air	0.007 m^3/s (15 ft^3/min)
Air leakage	3 ACH_{50}
Site	
Terrain	City
Ground reflectance	0.2

3.2. Climates, seasons, and weather

Simulations were conducted in cities representing four climate types that are widely represented in the U.S. and internationally, including humid subtropical (*Cfa*), humid continental (*Dfa/b*), semi-arid (*BSh/k*), desert (*BWh/k*), and Mediterranean (*Csb*), using weather files that represented typical meteorological conditions from 2004–2018 [65] (Table 3). All simulations were conducted during the month of May, chosen to provide conditions suitable for passive heating and cooling in many climates as well as a range of weather patterns: while May in continental and Mediterranean climates in the Northern hemisphere is often part of the residential heating season, the cooling season is well underway in humid subtropical and warm semi-arid climates [e.g. 16]. Additionally, many of these climates experience May outdoor air temperatures both warmer and cooler than the chosen thermostat setpoints, challenging the RL process to accommodate alternating heat gain and loss patterns. Dwelling microclimates were represented as urban, affecting local wind speed [66], and with low surface reflectance; site shading and urban heat island effects were not included (Table 2).

3.3. Learning architecture

A two-stage design strategy was employed for developing policies to control shading, movable insulation, and natural ventilation (Fig. 4). In

Table 3
Regions, climates, and weather files.

Region and city	Köppen[a]	IECC[b]	Weather file (TMYx.2004--2008.epw [65])
Pacific Coast			
Portland OR (training)	*Csb*	4C	USA_OR_Portland.Intl.AP.726980
Vancouver BC	*Csb*	4C	CAN_BC_Vancouver.Intl.AP.718920
Seattle WA	*Csb*	4C	USA_WA_Seattle-Tacoma.Intl.AP.727930
Eugene OR	*Csb*	4C	USA_OR_Eugene.AP-Sweet.Field.726930
Eureka CA	*Csb*	4C	USA_CA_Eureka-California.Redwood.Coast-Humboldt.County.AP.725945
San Francisco CA	*Csb*	3C	USA_CA_San.Francisco.Intl.AP.724940
West and Midwest			
Denver CO (training)	*BSk*	5B	USA_CO_Denver.Intl.AP.725650
Boise ID	*BSk/BWk*	5B	USA_ID_Boise.AP-Gowen.Field.ANGB.726810
Provo UT	*BSk/Dfb*	5B	USA_UT_Provo.Muni.AP.725724
Casper WY	*BSk/Dfb*	6B	USA_WY_Casper-Natrona.County.Intl.AP.725690
Omaha NE	*Dfa*	5A	USA_NE_Omaha-Eppley.AF.Intl.AP.725500
Wichita KS	*Dfa*	4A	USA_KS_Wichita.Eisenhower.Natl.AP.724500
Southeast			
San Antonio TX	*Cfa*	2A	USA_TX_San.Antonio-JB.San.Antonio-Randolph.AFB.722536
Houston TX	*Cfa*	2A	USA_TX_Houston-Bush.Intercontinental.AP.722430
Jackson MS	*Cfa*	3A	USA_MS_Jackson-Evers.Intl.AP.722350_TMYx.2004--2018.epw
Atlanta GA	*Cfa*	3A	USA_GA_Atlanta-Hartsfield-Jackson.Intl.AP.722190
Tallahassee FL	*Cfa*	2A	USA_FL_Tallahassee.Intl.AP.722140
North and Northeast			
Albany NY (training)	*Dfb*	5A	USA_NY_Albany.Intl.AP.725180
Bismarck ND	*Dfb*	6A	USA_ND_Bismarck.Muni.AP.727640
Milwaukee WI	*Dfa/Dfb*	6A	USA_WI_Milwaukee-Mitchell.Intl.AP.726400
Cleveland OH	*Dfa*	5A	USA_OH_Cleveland.Hopkins.Intl.AP.725240
Burlington VT	*Dfb*	6A	USA_VT_Burlington.Intl.AP.726170
Boston MA	*Dfa*	5A	USA_MA_Boston-Logan.Intl.AP.725090

[a]Köppen-Geiger climate designations: *BSk*: cold semi-arid (steppe); *BWk*: cold desert; *Cfa*: humid subtropical; *Csb*: Mediterranean; *Dfa*: hot-summer humid continental; *Dfb*: warm-summer humid continental [67].
[b]International Energy Conservation Code (IECC) climate zones: 2: Hot; 3: Warm; 4: Mixed; 5: Cool; 6: Cold; A: Humid; B: Dry; C: Marine [4].

the first stage, supervised learning was used to create an approximate representation of the expert policy π_E as a neural network (termed the *imitation network*) (Fig. 4a). This approximate representation (π_θ) was given the same parameterization as the RL agent to be trained, allowing it to be used to initialize the learning process. In the second stage, this initial guess was refined via RL to create distinct climate-adapted agents (Fig. 4b).

3.3.1. Stage 1: Imitation learning from the expert

To create an imitation of the expert policy π_E, we first parameterized a generic policy π by a set of basis functions *identical* in structure to the policy used in the RL training and by a parameter vector θ that was identified to best mimic the expert. To determine this optimal parameter vector, we first collected observation and expert action pairs, $\{y_i, \pi_E(y_i)\}$, respectively. The parameter vector θ_0 was then determined by solving:

$$\theta_0 = \arg\min_{\theta} d(\pi_E(y_i), \pi(y_i, \theta), i \in \{1, 2, 3 \cdots\}), \tag{5}$$

where $d(\cdot,\cdot)$ is an appropriate distance metric in the action space. The parameterized policy π_{θ_0} is therefore the best approximation of the expert, evaluated by the metric d.

The expert demonstration dataset was collected by simulating the model dwelling (Section 3.1) under control of the expert policy, in 8 cities representing the four climates of interest (Fig. 5; Table 3), to yield a generalized representation of the expert's decisions. Each demonstration sample was stored as a state–action pair $(s_{(m,t)}, \pi_E(s_{(m,t)}))$, in which $s_{(m,t)}$ denotes the observation (state) of city m at time t, and $\pi_E(s_{(m,t)})$ indicates the corresponding decision by the expert policy π_E. The aggregated dataset therefore consisted of $D = \{(s_{(m,t)}, \pi_E(s_{(m,t)}))| \ \forall (m,t) \in \mathcal{M} \times \mathcal{T}\}$, where $\mathcal{M}$ indicates the set of cities and $\mathcal{T}$ indicates the set of timesteps in which the state–action pairs were recorded. Simulation in each the 8 cities over the single month of May, in 15-min timesteps, yielded approximately 23,000 expert demonstration state–action pairs. Dataset D was then divided into a training dataset and a

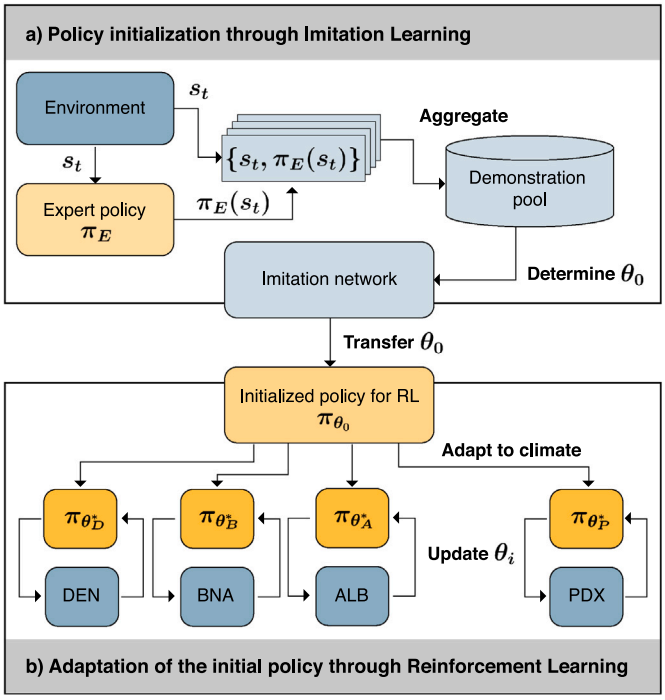


Fig. 4. Development of climate-adapted RL controllers. (a) Simulations of the study dwelling in eight cities with contrasting climates (Fig. 5), under control of the expert policy π_E (Table 1), generated a demonstration pool of state (s_t) and action ($\pi_E(s_t)$) pairs. This demonstration pool was then used in supervised learning to train the imitation network, θ_0. (b) Imitation model parameters were next imported into an RL framework to create the initial policy, π_{θ_0}. Subsequent training, using the policy-gradient REINFORCE algorithm, generated policies π^*_θ adapted to the cold semi-arid climate of Denver CO (DEN); humid subtropical Nashville TN (BNA); continental Albany NY (ALB); and Mediterranean Portland OR (PDX), respectively (Table 3).

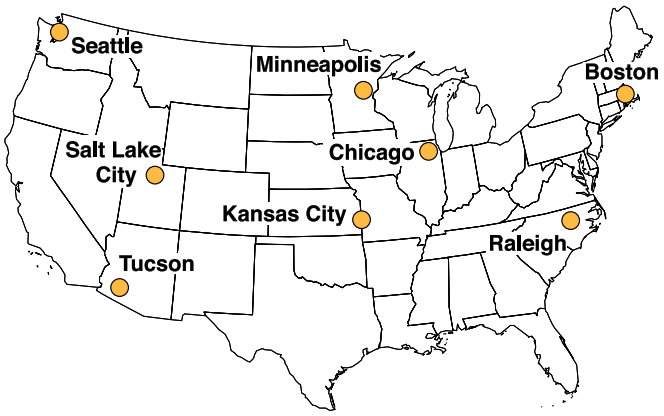


Fig. 5. Cities simulated under control of the expert policy to generate data for imitation learning. Cities were chosen to represent contrasting, globally significant climates: Mediterranean (*Csb*; Seattle WA); cold semi-arid (*BSk*; Salt Lake City UT); hot desert (*BWh*; Tucson AZ); hot-summer continental (*Dfa*; Chicago IL; Boston MA); warm-summer continental (*Dfa/b*; Minneapolis MN); and humid subtropical (*Cfa*; Kansas City MO; Raleigh NC).

Fig. 6. Cities trained by RL for climate adaptation. Large, black-bordered dots indicate training cities representing the Pacific Coast (teal), West and Midwest (orange), North and Northeast (dark blue), and Southeast (light blue) regions and corresponding climates; small dots indicate cities in which the resulting policies were tested. See Table 3 for weather file details.

test dataset in a ratio of 4:1; the training dataset was used to train π_θ as a classifier (Section 2.4), while the test dataset was used to evaluate the accuracy of the imitation network in reproducing the expert policy (Section 4.1).

3.3.2. Stage 2: Climate-specific policy refinement

Once the parameter vector θ_0 was optimized so that the imitation network matched the expert policy as closely as possible, these learned parameters were used as the initial guess for the policy parameters in subsequent learning. These parameters were then trained through RL in semi-arid Denver CO, humid subtropical Nashville TN, continental Albany NY, and Mediterranean Portland OR (Table 3; Fig. 6). In each climate, parameters were updated in the direction that maximized the sum of expected future rewards, allowing the policy to learn climate-adapted strategies for shading, natural ventilation, and in some cases, movable insulation control.

3.3.3. RL algorithm

For compatibility, the definitions of the states s and actions a for the RL MDP were chosen to be identical to those of the expert policy, as discussed in Section 2.4, i.e., $s = s_t$ and $a = \pi_E(s_t)$. Next, the reward function $r(\cdot)$ was designed to maximize utilization of passive resources for the reduction of sensible heating and cooling loads. (Latent loads were negligible under the conditions investigated; even in humid subtropical cities, peak outside air temperatures coincided with relative humidities of 55% or less, and the effectiveness of natural ventilation in removing moisture generated by occupancy kept indoor relative humidities below levels that would have compromised thermal comfort [68].) The reward was therefore defined as the weighted sum of r_1 and r_2, where r_1 incentivized reduction of the total (sum of heating and cooling) load at time t, and r_2 was associated with the cost of actuation for each passive element, as shown below,

$$r(E_t, a_t, a_{t-1}) = w_1 r_1 + w_2 r_2, \quad \text{where} \tag{6}$$

$$r_1 = \begin{cases} +1 & E_t = 0 \\ -1 - E_t/C & E_t > 0, \end{cases} \quad \text{and}$$

$$r_2 = \begin{cases} +1 & a_t = a_{t-1} \\ -1 & a_t \neq a_{t-1}. \end{cases}$$

We note that the policies π_θ for both imitation and reinforcement learning algorithms were given *identical parameterization*, though they may have had different parameters θ. Although these policies were functionally parameterized by fully connected neural networks, the parameterized policy was structured such that the flow of information mimicked the expert policy (Table 1). Specifically, two separate networks were used to generate the actions corresponding to shading and natural ventilation control (Fig. 7). The rationale for this choice was two-fold: first, directing the necessary information to each network separately improved the imitation accuracy, and second, the ability to freeze or further train individual networks provided the freedom to manage each controller and the associated network separately.

Therefore, each observation s_t was split into two parts: components corresponding to indoor air temperature and window heat flux histories were used to generate shading decisions, while indoor and outdoor air temperature histories were used to generate natural ventilation decisions (Fig. 7). Each input observation component was fed to the input layer of a separate neural network, each consisting of three hidden layers with 100 nodes. Rectified Linear Unit (ReLU) functions were chosen as the activation functions for each individual neuron in the network. The output layers of the shading and ventilation neural networks were designed with 4 and 7 nodes, respectively, with each output node mapping to a specific action as shown in Table 1. Finally, the values at each of these output nodes represented the *probabilities* (e.g., p(shading = on), p(shading = off), as shown in Fig. 7), that a particular action would be taken. Thus, the policy π mapped observations s_t to probability distributions of actions $p(A_t = a_t)$. Since the expert policy π_E was deterministic, the corresponding probability distribution collapsed to the expert action, setting the probabilities of all other actions to zero.

3.3.4. RL training hyperparameters and software implementation

The values of the learning rate α for imitation learning and for RL were chosen to be 10^{-4} and 5×10^{-3}, respectively, with a lower learning rate assigned to RL to limit the loss of expert knowledge gained during initialization. The discount factor, γ, was heuristically selected to be 0.99. Algorithm implementation and validation were accomplished with Python v3.6, MATLAB 2021a, and EnergyPlus v9.2. Python was used to implement the policy network, generate actions, and compute gradients, while MATLAB was used as a communication intermediary between EnergyPlus and Python, supported by the MATLAB-EnergyPlus Co-simulation Toolbox (MLEP) [69]. The high-level decisions made in Python, i.e., the action decisions given by a

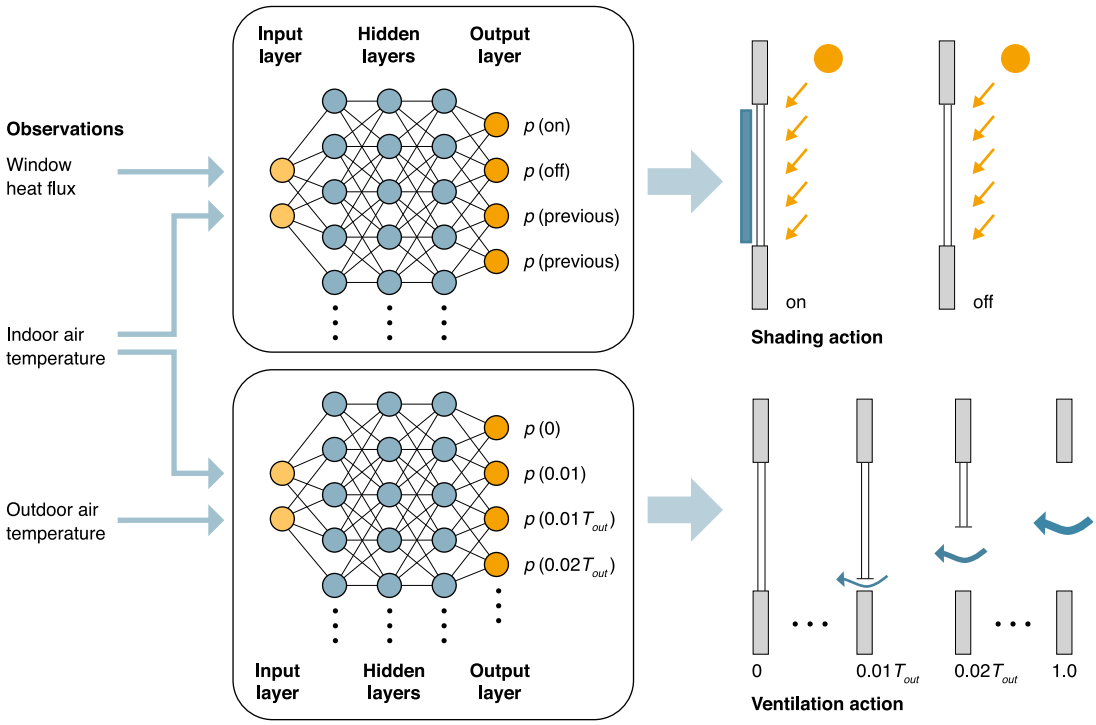


Fig. 7. Structure of the neural networks used for policy parameterization. Two neural networks, one dedicated to shading and one to natural ventilation, worked in tandem to determine the probability that each action would be taken. The policy was structured such that the flow of information reproduced that of the expert policy, in which observations from the environment were split and transformed into input forms for each network. Each neural network had three hidden layers, each with 100 nodes; nodes of the output layer corresponded to the probability that each possible shading and ventilation control action would be taken, respectively.

policy, were communicated to MATLAB through a TCP/IP port. Each action was translated into a specific schedule command and deployed in the EnergyPlus co-simulation through MLEP; observations acquired from the EnergyPlus simulation were returned to Python through the same TCP/IP port. The RL algorithm (REINFORCE) was implemented in Python, using TensorFlow for all neural network-related computations. Since the imitation learning did not require communication between EnergyPlus and Python, the static training dataset for imitation learning was generated in MATLAB and then loaded in Python for training. All simulations and training algorithms were executed on an NVIDIA Quadro P4000 GPU and an Intel i7-9700K CPU. Imitation learning was complete within 10,000 training iterations and 2 min, while reinforcement learning required up to 900 iterations and 8 h. Although the RL training process involved fewer iterations, each EnergyPlus simulation of the model dwelling over one month, i.e. each single iteration, required approximately 35 s.

3.4. Study scope and limitations

For brevity, clarity, and depth in the results, and to support the central purpose of evaluating the potential of a new method, the scope of this study was limited in several ways. First, it addressed only a single space type, that of a one-bedroom apartment with a single exterior exposure. This type was chosen for its abundance throughout the U.S. and the world; at the same time, the results described below may not be directly applicable to other residential spaces or to non-residential buildings. For analogous reasons, the time period of investigation was limited to a single month. This interval allowed consistent weather patterns to be experienced within each city, facilitating clear and detailed analyses of imitation learning accuracy and of policy changes made by climate-adaptive reinforcement learning. Investigation of season-to-season adaptations will be an important direction for future work, however. Additionally, for comparability among the range of climates investigated, the reward structure valued only the sum of sensible heating and cooling loads. In other words, rewards did not consider variations in the relative efficiencies of specific mechanical heating and cooling systems, nor in regional differences in the fuel mixes used to power them, although these are known to affect the relative energy consumption and carbon emission intensities of space cooling and heating. Policy actions therefore minimized space conditioning loads rather than energy consumption, carbon emissions, or cost. The reward construction also omitted the influence of relative humidity on thermal comfort, since this was negligible in the month chosen (see Section 3.3.3), although its inclusion will be important under warmer and more consistently humid conditions. Finally, RL training excluded rewards for providing daylight and/or views, to allow RL policy actions to be interpreted without ambiguity; additionally, residential spaces are less occupied than workplaces during daytime hours, when solar heat gain occurs, even when occupants are present. Still, adding daylighting and view criteria to RL reward structures will be a valuable advance, particularly for the integration of passive cooling and heating systems into workplaces.

4. Results and discussion

4.1. Imitation network accuracy and performance

The accuracy of the imitation network π_{θ_0} in mimicking the expert policy was first evaluated with a subset of data from the demonstration pool, representing 20% of the total, that had been reserved for testing (see also Section 3.3.1). In this evaluation, 1000 of the control decisions made by the imitation network were compared to those that would have been made by the expert policy; results are shown in confusion matrices, which array predicted (i.e., imitation) values in one dimension and the corresponding true (i.e., expert) values in another. The fraction of correct predictions therefore appears along a diagonal series of cells, with fractions of incorrect predictions shown in the off-diagonal positions (Fig. 8).

The imitation network correctly emulated the expert policy in the great majority ($\geq$95%) of shading actions (Fig. 8a); in the remainder,

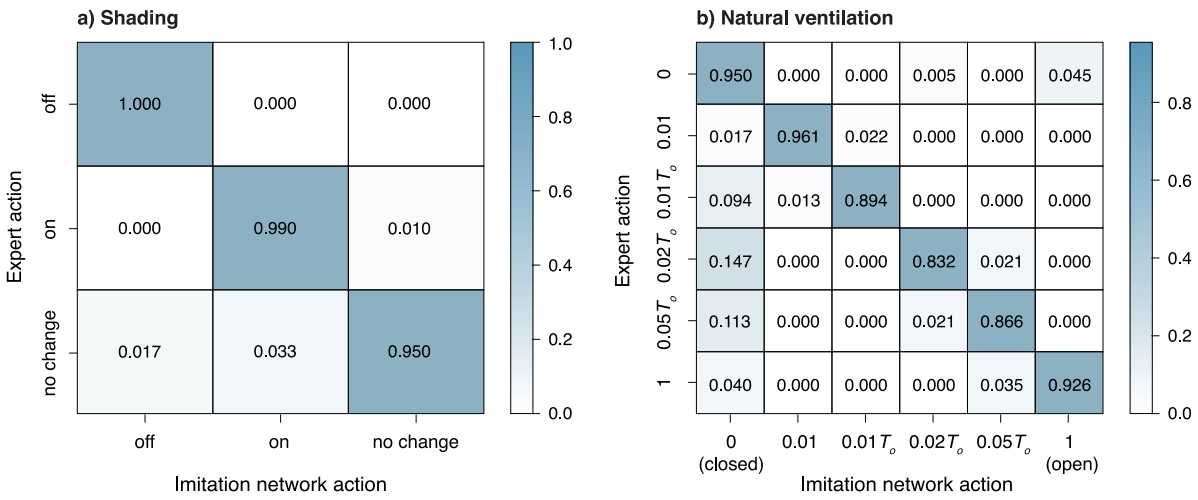


Fig. 8. Accuracy of the imitation network in reproducing (a) shading and (b) natural ventilation control actions of the expert policy. Actions taken by the imitation network in 1000 samples of the imitation testing dataset (Fig. 5; Section 3.3.1) are shown horizontally; expert actions that would have been taken under the same conditions are shown vertically. Values show the fraction of each imitation network action that corresponded to each expert action.

the imitation primarily activated shading under conditions in which the expert would have maintained the previous position, slightly increasing the shading frequency relative to the expert policy. Its accuracy in reproducing the expert ventilation actions was somewhat lower (> 83%), a finding partly explained by the inclusion of finer gradations in the ventilation actions. Additionally, in many discrepancies, the imitation network chose actions only one step removed from the action the expert would have chosen. The majority of discrepancies occurred in the imitation network's choice of the 'closed', or 0, action, which occurred in up to 15% of the cases in which the expert would have chosen a partial aperture opening. In other words, the imitation network somewhat reduced natural ventilation availability relative to the expert policy.

We next investigated the effectiveness of the imitation network in reducing space conditioning loads in 24 cities representing semi-arid, humid subtropical, continental, and Mediterranean climate types (Table 3), grouped geographically into four U.S. regions (Fig. 6). With the exception of Seattle and Boston, these cities had not been included in the imitation training or testing datasets (see also Section 3.3.1). Because the state distribution encountered during this evaluation was different from the distribution in the training data, however, the performance of π_{θ_0} depended exclusively on the effectiveness of the learned policy. Baseline dwellings, i.e. those without passive systems, experienced sensible cooling loads in May ranging from approximately 650 MJ along the Mediterranean Pacific Coast to nearly 1600 MJ in the humid subtropical Southeast, reflecting each region's combination of window solar heat gain; internal heat gain from occupancy, lighting, and equipment; and envelope heat loss (Fig. 9). The latter reflected the influence of contemporary (2018) building code requirements [4]; variations in building heat gain and loss properties are explored, however, in Section 4.5. No space heating loads were observed under these conditions.

Use of the imitation network to control shading and natural ventilation in these dwellings reduced May loads substantially. In cities of the West and Midwest, with semi-arid and continental climates, baseline loads were reduced by over 80%, to levels of 120 MJ or less. Notably, the use of natural ventilation to reduce cooling loads also induced small heating loads in this region (Fig. 9a), illustrating an area for improvement through RL. Similarly, in the continental climates of the Northern cities (Fig. 9c), and in the Mediterranean climates of the Pacific Coast cities (Fig. 9d), imitation policies reduced total loads by 70% or more, often with the induction of heating loads from imprecise natural ventilation control. In the humid subtropical cities of the Southeast, cooling loads were also reduced considerably, but to lesser extents. In Houston, for example, the cooling load decreased by only 47%, remaining near 800 MJ (9b). Such differences in performance among climates were expected: late spring is warmer in humid subtropical areas than in the other climates shown [65], creating higher cooling loads; further, humid climates often experience warmer nighttime temperatures, limiting the utility of natural ventilation [15]. Still, these results showed the strong potential of well-controlled, well-integrated shading and natural ventilation to reduce heating and cooling loads across a broad range of climate types, including the humid subtropics. At the same time, the new heating loads that appeared above, as well as the excellent performance of optimized, climate-specific controls for integrated shading and natural ventilation systems [e.g. 17,22,36], suggested that further potential remained.

4.2. Climate adaptation of actions through reinforcement learning

We next sought to improve the imitation network through RL training in semi-arid Denver CO, humid subtropical Nashville TN, continental Albany NY, and Mediterranean Portland OR (Table 3) to yield climate-adapted policies. The imitation network π_{θ_0} was used as the initial policy in each case; θ was updated at the end of each training episode, using the REINFORCE algorithm, until converging to θ^* (Fig. 4b). Each RL training episode required fewer than 900 iterations to converge, showing notable improvement over previous RL methods, unassisted by imitation learning, which required up to 1700 iterations [38].

RL training modified the imitation policies in noticeable, consistent ways that corresponded to the training cities' contrasting needs and climatic resources, as shown by comparing control actions, indoor air temperatures, and heating and cooling loads between dwellings controlled by the imitation network and those controlled by the climate-adapted RL policies (Fig. 10). In Albany, for example, May is a relatively cool month, and control of passive systems by the imitation policy created new heating loads through its use of natural ventilation. In response, the RL policy added movable insulation to the windows on most nights, reducing nighttime window heat loss, and reduced both the frequency of natural ventilation and the extent of ventilation aperture opening. As a result, the dwelling's indoor air was kept about 1.5 °C warmer under control of the RL policy than under control of the imitation policy, virtually eliminating the new space heating loads without adding

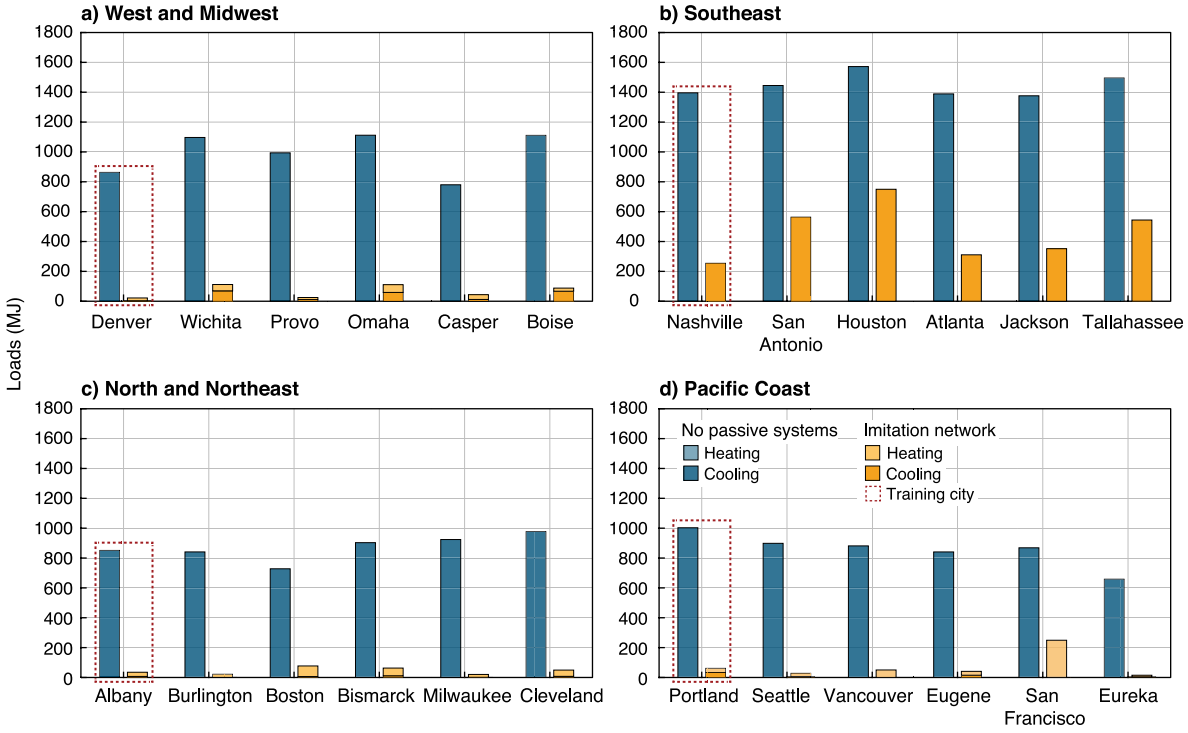


Fig. 9. Performance of the imitation network, π_{θ_0}, in reducing residential space heating and cooling loads during the month of May. Loads are compared in the absence of passive systems (blue) and in the presence of operable shading/movable insulation and natural ventilation controlled by the imitation network (orange) in four U.S. regions: (a) the West and Midwest (semi-arid and continental); (b) the Southeast (humid subtropical); (c) the North and Northeast (continental); and (d) the Pacific Coast (Mediterranean). Cites used in subsequent reinforcement learning, described below, are indicated in red dotted boxes.

cooling loads (Fig. 10a). The final (trained) RL policy maintained the daytime shading actions of the imitation policy, however, both because solar heat gain was not needed for warmth and because daylight and/or view access were not rewarded. The latter considerations were beyond the current scope (Section 3.4), but they could readily be included in future work.

In humid subtropical Nashville, in contrast, the climate-adapted RL policy adjusted the imitation policy's daytime shading policy only slightly (Fig. 10b). Instead, it increased natural ventilation substantially, maintaining nighttime indoor air temperatures 2–3 ℃ below those found in the dwelling controlled by the imitation policy. By cooling the floor and other materials in the space, as well as the air, this behavior allowed the dwelling to reach the upper thermostat setpoint later in the day than it otherwise would have. As a result, cooling loads were reduced appreciably (quantified below). Notably, further passive cooling capacity existed in the form of cool night air than the RL policy was able to use, due to the imposition of a lower thermostat setpoint (18 ℃). If this setpoint had been lowered further, or even eliminated in unoccupied rooms, the residual loads could have been reduced even further.

4.3. Portability of RL policies among related climates

We next investigated the abilities of the four climate-adapted RL policies to reduce total space conditioning loads in related yet unfamiliar climates: if policy performance remained comparable among cities with related climates, the need for time-consuming (re-)training would be greatly reduced, facilitating general application.

Comparison of total space conditioning loads between each of the RL training cities (Denver, Nashville, Albany, and Portland) and unfamiliar cities in their respective climate types (Fig. 6; Table 3) showed that control by the climate-adapted RL policies reduced loads to levels below those achieved by the imitation policy in virtually all cases. Additionally, in most of the unfamiliar cities, RL control reduced loads to comparable or greater extents than in the training cities (Fig. 11). In the semi-arid and continental West and Midwest, for example, the Denver-trained RL policy reduced the loads that remained, under control of the imitation network, by 16%–48% in Wichita, Provo, Omaha, Casper, and Boise, compared to a reduction of 40% in Denver itself (Fig. 11a). The remaining loads totaled approximately 80 MJ or less in each case, representing reductions from the original (shown in Fig. 9a) of 93% or more.

In the humid subtropical Southeast, the Nashville-trained RL policy was also able to improve upon the imitation policy, but to a lesser extent, reducing the remaining loads by 5%–15% in San Antonio, Houston, Jackson, and Tallahassee, compared to a reduction of 23% in Nashville itself (Fig. 11b). In part, this reflected the ability of the imitation policy to operate the passive systems effectively in this climate, given the cooling resources available, leaving little room for improvement. In Atlanta, however, the changes to shading and natural ventilation control made by the climate-adapted RL policy, analogous to those shown for Nashville in Fig. 10b, caused total loads to rise by about 6%. This increase resulted primarily from Atlanta's greater May solar radiation intensity (Table 3), increasing the dwelling's solar heat gain and inducing greater use of natural ventilation. Greater natural ventilation in this unfamiliar climate, in turn, replaced a small fraction of the cooling load with a slightly greater heating load. Still, the RL policy reduced initial loads (shown in Fig. 9b) by 65% or more in all of the humid subtropical cities (including Atlanta) except Houston, in which it nevertheless reduced initial loads by more than 50%.

In the continental North and Northeast, the Albany-trained RL policy reduced the combination of cooling and heating loads that remained, under imitation network control, by 65%–90% in Burlington, Boston, Milwaukee, and Cleveland, compared to approx. 70% in Albany itself (Fig. 10c). As a result, RL control nearly eliminated May space conditioning loads in four of the five unfamiliar cities, achieving reductions of 98% or more; this was accomplished by adding movable insulation to windows at night, reducing heat loss, and by lessening

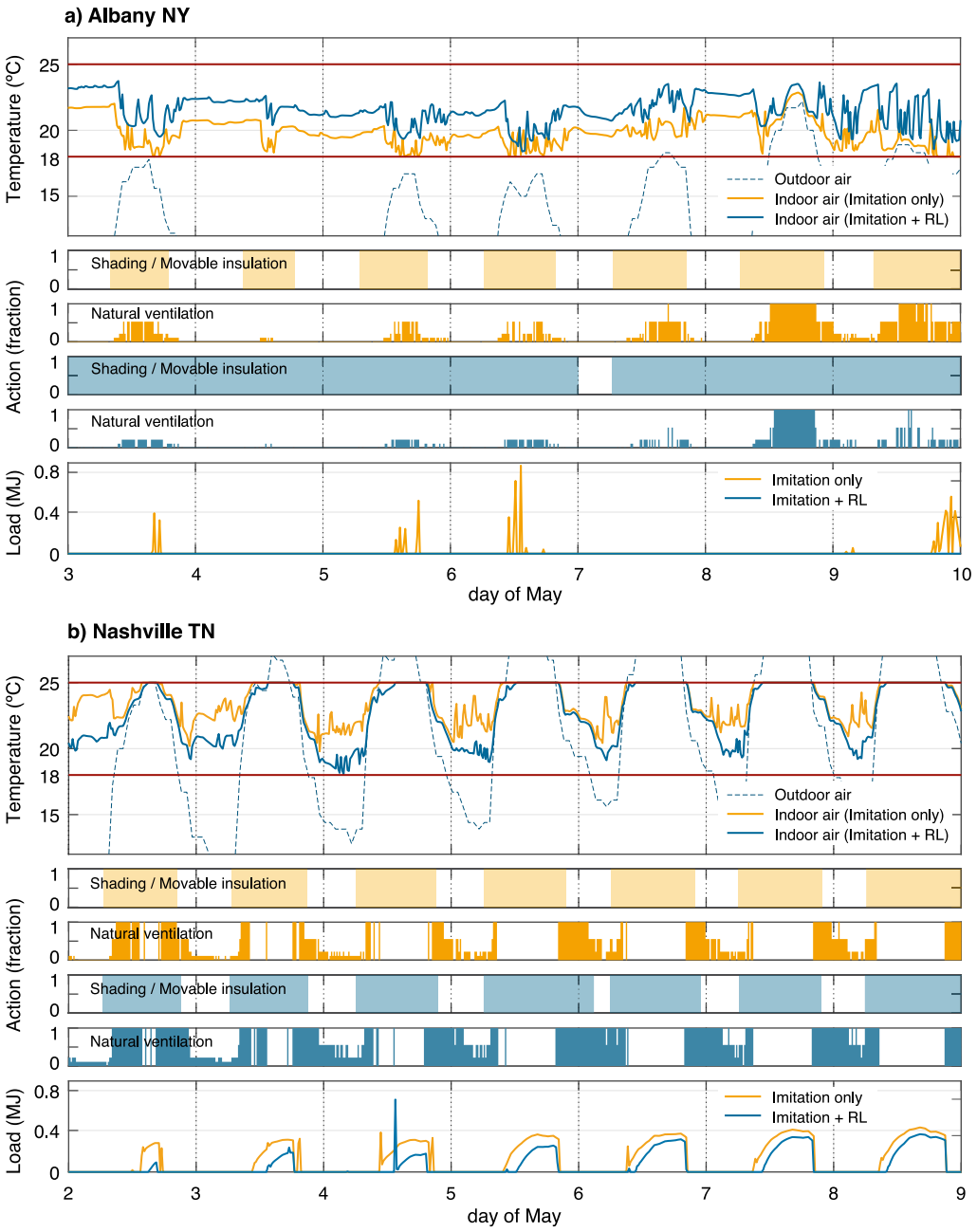


Fig. 10. Climate adaptation of control policies through RL. Indoor and outdoor air temperatures, shading/movable insulation and natural ventilation control actions, and sums of heating and cooling loads are shown for the study dwelling under control of imitation network policies (yellow) and of climate-adapted RL policies (blue), simulated in (a) Albany NY (continental, *Dfb*) and (b) Nashville TN (humid subtropical, *Cfa*) over representative weeks of May; weather data were provided by TMYx 2004–2018 weather files (Table 3).

natural ventilation. Even in Bismarck, with the highest final load in the region of 35 MJ, the initial load (shown in Fig. 9c) was reduced by over 95% by the climate-adapted RL policy. In the Mediterranean Pacific Coast region, finally, the Portland-trained RL policy reduced the remaining loads by 33%–88% among the cities of Seattle, Vancouver, Eugene, San Francisco, and Eureka, although it reduced loads in Portland itself by only 16% (Fig. 10d). In each of these cities, however, the Portland-trained RL policy reduced initial loads (shown in Fig. 9d) by over 90%, echoing the findings in the Western and Northern regions.

These results show that imitation-assisted policy-gradient RL is highly effective in generating policies that control shading, movable insulation, and natural ventilation simultaneously, reducing space conditioning loads dramatically when passive cooling and heating resources are present. Integrated control of passive cooling and heating elements has been investigated only rarely [e.g. 17,52], and the finding that REINFORCE can generate multiple effective climate-adapted policies for this challenging problem is a notable contribution. Of greater significance, these results also show for the first time that climate-adapted policy gradient RL policies are highly portable among cities within related climates, suggesting that far less climate-specific learning may be required than previously believed.

4.4. Advantage provided by use of the expert system

To reveal the performance advantage provided by initialization with the expert system, we next trained RL agents without it, in each of the four training climates (Denver, Nashville, Albany, and Portland), in two ways. One set of agents was trained with the two-part neural network used above (Fig. 7), in which policies for shading and natural ventilation actions were learned separately. Although this structure was

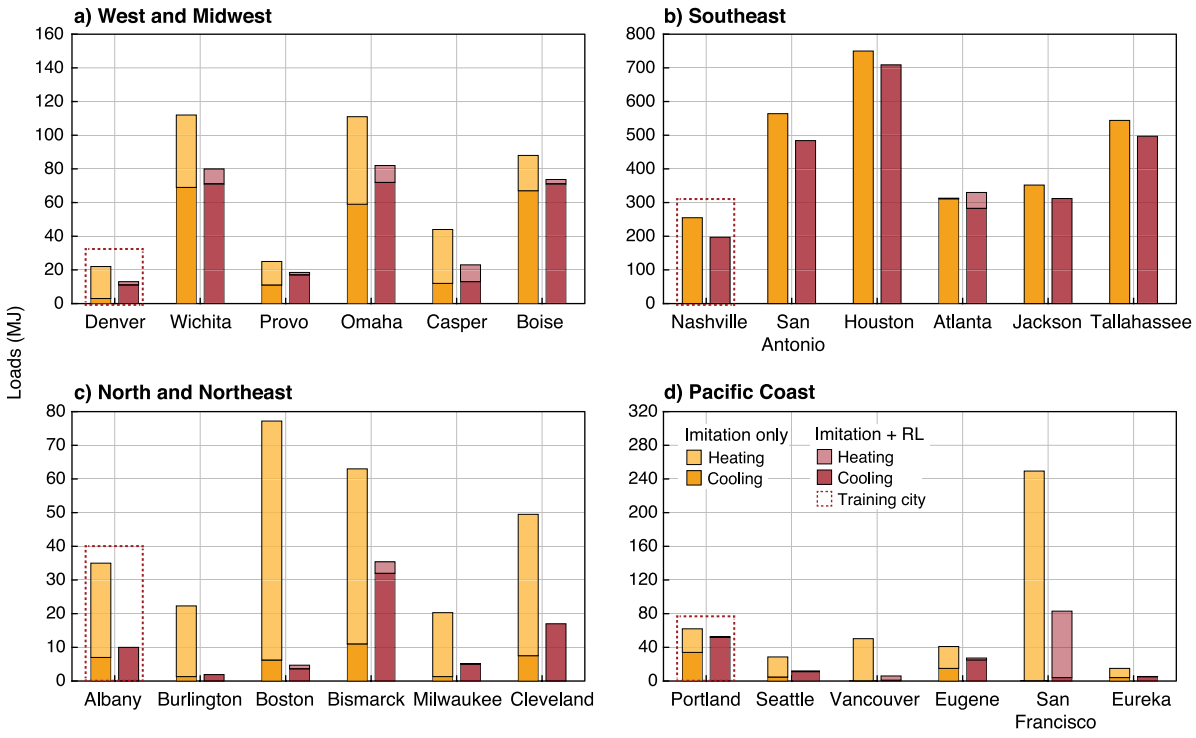


Fig. 11. Performance of climate-adapted RL control policies. Heating and cooling loads represent totals over the month of May in simulations of the study dwelling under control of the imitation network (orange, consistent with Fig. 9) and under control of trained RL policies (red), in (a) the West and Midwest (semi-arid and continental), (b) Southeast (humid subtropical), (c) North and Northeast (continental), and (d) Pacific Coast (Mediterranean) regions. Climate-adapted policies were initialized through imitation learning and adapted further through reinforcement learning in the training cities, shown in red dotted boxes (see also Fig. 4).

appropriate for RL following imitation learning, as described above (Section 3.3.3), it limited the ability of a naïve agent (i.e., one without the knowledge of an expert system) to coordinate shading and natural ventilation control. Because of this, another set of agents was trained with a single neural network, integrating shading and natural ventilation control actions, analogous to that of Park et al. [38].

Comparison of total space conditioning loads among dwellings controlled by each of the three RL agents showed, strikingly, that the agents pre-trained through imitation learning achieved the greatest load reductions by substantial margins in each region (Fig. 12). In the semi-arid and continental West and Midwest, for example, pre-trained agents reduced loads by 70% or more, compared to the agents trained by RL alone, in 5 of the 6 cities examined. In the sixth, the pre-trained agent reduced loads by about 25%. Pre-trained agents also reduced loads by 70% or more in the humid subtropical Southeast, compared to RL-only agents. In this region, however, control by RL-only agents *increased* baseline cooling and heating loads in several cities (compare Figs. 12b to 9b), showing that pre-training also contributed to agent portability among related climates. Likewise, in the North and Northeast (continental) and Pacific Coast (Mediterranean) regions, pre-trained agents reduced loads by 70%–98%, compared to those achieved by agents trained by RL alone, with only one exception. In that exception, occurring in oceanic San Francisco, the pre-trained agent nevertheless improved slightly upon the RL-only agent; in no case did pre-trained agents perform worse than agents trained by RL alone. Together, these results confirm that the performance of RL agents in shading and natural ventilation control can be improved markedly by imitation learning.

4.5. Robustness to variation in dwelling thermal parameters

We next sought to understand the robustness of the climate-adapted RL policies to common variations among apartment dwellings within a single climate, focusing on parameters known to affect residential heating and cooling loads appreciably: exterior facade orientation; window glazing assemblies; heat gain rates from people, lighting, and equipment; and infiltration rates. Results are illustrated in two cities with contrasting climates: continental Burlington VT, in which typical May outdoor air is frequently cooler than the lower thermostat setpoint (18 °C), and humid subtropical Houston TX, in which May outdoor air is often warmer than the upper thermostat setpoint (25 °C); these cities are comparable climatically to Albany NY and Nashville TN, respectively, shown in Fig. 10. Within these climates, RL policy performance was compared among (i) exterior orientations of North, East, and West, in comparison to the original orientation of South; (ii) window assemblies representing double-paned clear glass and triple-paned low-emissivity (LowE) glass, in comparison to the original double-paned LowE assembly; (iii) internal heat gain rates of 0.75×, 1.0×, and 1.25× the original rate of 9.15 W/m^2; and (iv) air leakage rates of 1.0× and 1.67× the original value of 3 ACH_{50}.

4.5.1. Robustness to variation in orientation

In both continental Burlington and subtropical Houston, dwelling orientation affected baseline cooling loads noticeably: dwellings with East-facing windows showed the highest loads, resulting from morning solar heat gain retained throughout the day, while Western exposures experienced the second-greatest loads, showing the influence of window heat gain from low-altitude afternoon sun; these exposures both had higher cooling loads than the original South-facing dwelling. In contrast, North-facing dwellings experienced lower cooling loads than the original, as expected, since their exposure to direct sun was lowest.

In Burlington, the continental climate-adapted RL policy (trained in Albany NY; Fig. 6) reduced the baseline cooling load in the original South-facing orientation by over 99% (Table 4; see also Figs. 9c and 11c). Remarkably, this RL policy performed just as well in North, East, and West-facing dwellings, again reducing loads by over 99%

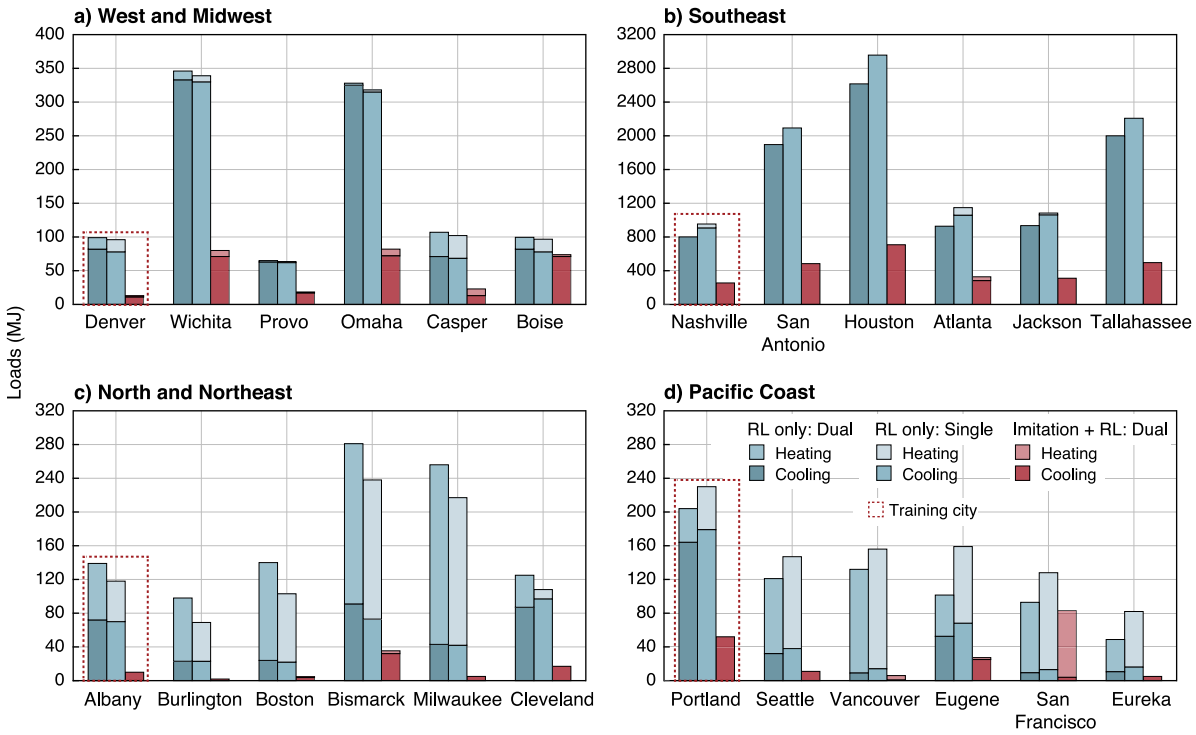


Fig. 12. Performance comparison among RL agents trained with and without pre-training by the imitation network. Heating and cooling loads represent totals over the month of May in simulations of the study dwelling under control of each agent type. Training by RL alone used either the dual neural network above ("RL only: Dual") or a single neural network combining shading and natural ventilation actions ("RL only: Single"). RL training following initialization with the imitation network used the dual network above ("Imitation + RL: Dual", in red, consistent with Fig. 11). Results are shown for (a) West and Midwest (semi-arid and continental), (b) Southeast (humid subtropical), (c) North and Northeast (continental), and (d) Pacific Coast (Mediterranean) regions. Cities in which regional training occurred are indicated by red dotted boxes.

in each case. Analogous results were found in Houston: in this city, the humid subtropical climate-adapted RL policy (trained in Nashville TN; Fig. 6) reduced the baseline cooling load of the original South-facing dwelling by approximately 55%, whereas it reduced the cooling loads of North, East, and West-facing dwellings by 53% or more. The climate-adapted RL policies therefore maintained greater performance consistency than expected throughout the range of possible dwelling orientations, including the demanding East and West exposures, showing that dwellings of all orientations within a multifamily building have the potential to benefit comparably from such control of passive space-conditioning systems.

4.5.2. Robustness to variation in glazing assembly

The substitution of the original double LowE glazing assembly with a double clear assembly, raising both the thermal transmittance and solar heat gain coefficient (Table 4), caused minimal changes to the original cooling loads: in Burlington, the cooling load was reduced by about 1%, while in Houston it rose by about 5%, reflecting the near balance in May between the increased solar heat gain and the increased window heat loss that resulted. Accordingly, the climate-adapted RL policies performed as well in dwellings with double clear assemblies as in those with the original double LowE assemblies, reducing cooling loads by over 99% in Burlington and by over 57% in Houston. Substitution of the original windows with triple LowE assemblies, in contrast, lowered baseline cooling loads by 15% in Burlington and by 10% in Houston, reflecting their lower solar heat gain coefficients and thermal transmittance values. Again, however, the respective RL policies maintained the levels of performance observed in the original dwelling: in Burlington, the baseline cooling load was reduced by over 99%, while in Houston, it was reduced by nearly 50%.

4.5.3. Robustness to variation in internal heat gain

Variation in internal heat gain rates, representing variation in occupancy patterns, lighting intensity, and use of electrical equipment (e.g. appliances), affected baseline cooling loads dramatically in both cities, as expected. Reduction of the original internal gain rate of 9.15 W/m^2 [4] by one-quarter lowered cooling loads by about 30% in Burlington and by about 18% in Houston; similarly, raising internal gains by one-quarter increased cooling loads by about 30% in Burlington and about 17% in Houston, revealing the relatively greater contribution of internal heat gain to cooling loads in Burlington's cooler continental climate (Table 4). As above, however, the climate-adapted RL policies performed comparably over the range of internal heat gain levels. In Burlington, the continental RL policy reduced cooling loads by over 99% in all three cases, allowing only slight (<1%) increases at the unfamiliar heat gain levels. In Houston, similarly, the humid subtropical RL policy reduced the cooling load by about 52% in the dwelling with the higher internal heat gain rate, compared to about 55% in the original dwelling; however, it reduced the load by over 59% in the dwelling with the lower internal gain rate. This shows that the RL policy was able to accommodate the lower internal heat gain to the advantage of cooling, without overcooling (which remains possible in May even in the humid subtropics; Fig. 10b), giving it the potential to adapt to increasing efficiency in lighting and electrical equipment in a space over time.

4.5.4. Robustness to variation in air leakage

Variation in air leakage is a fourth important contributor to space conditioning loads in buildings, as well as a complicating factor in natural ventilation control. Air leakage rates reflect wall and window sealing methods, and older buildings typically have higher infiltration and exfiltration rates than newer ones. Increasing air leakage from the original value of 3 ACH_{50} to 5 ACH_{50} lowered the baseline cooling load in Burlington by about 10%, as expected given its cool May air, but raised it slightly (by <1%) in Houston due to its relative warmth. Still, consistent with the results above, the continental RL policy reduced the cooling load in the unfamiliar 5 ACH_{50} dwelling in Burlington by over

Table 4
Robustness of final RL policies to dwelling variations.

Orientation	Glazing assembly	Internal gains (W/m^2)	Air leakage (ACH_{50})	Cooling load[a]: Baseline (MJ)	Cooling load[b]: Trained policy (MJ)	Load reduction (%)
Burlington VT (continental)						
South	Dbl LowE[c]	9.15	3	841.4	1.7	99.8
North	Dbl LowE	9.15	3	565.6	0.4	99.9
East	Dbl LowE	9.15	3	1088.4	4.4	99.6
West	Dbl LowE	9.15	3	903.8	1.5	99.8
South	Dbl Clr[d]	9.15	3	830.9	2.6	99.7
South	Tpl LowE[e]	9.15	3	713.7	5.6	99.2
South	Dbl LowE	4.57 (0.75×)	3	574.2	5.2	99.1
South	Dbl LowE	18.3 (1.25×)	3	1103.1	9.4	99.1
South	Dbl LowE	9.15	5 (1.67×)	759.4	4.1	99.5
Houston TX (humid subtropical)						
South	Dbl LowE	9.15	3	1572.2	709.5	54.9
North	Dbl LowE	9.15	3	1534.1	707.3	53.9
East	Dbl LowE	9.15	3	2080.9	776.9	63.7
West	Dbl LowE	9.15	3	1841.4	734.3	60.1
South	Dbl Clr	9.15	3	1653.0	699.2	57.4
South	Tpl LowE	9.15	3	1417.7	718.4	49.3
South	Dbl LowE	4.57 (0.75×)	3	1297.7	529.2	59.2
South	Dbl LowE	18.3 (1.25×)	3	1835.4	879.1	52.1
South	Dbl LowE	9.15	5 (1.67×)	1577.5	721.2	54.3

[a]No heating loads were observed in these cases.
[b]No heating loads were observed in these cases.
[c]Double LowE Argon: U = 1.61 W/m^2K, SHGC = 0.55.
[d]Double Clear Air: U = 2.70 W/m^2K, SHGC = 0.70.
[e]Triple LowE Argon: U = 0.70 W/m^2K, SHGC = 0.30.

99%; similarly, the humid subtropical RL policy reduced the cooling load in the 5 ACH_{50} Houston dwelling by over 54%, a value almost identical to that observed in the original dwelling. Climate-adapted RL policies were therefore able to accommodate variations in air leakage as well as they had responded to variations in orientation, window glazing, and internal heat gain, further supporting the implication that such policies have the potential to perform well across a range of dwelling variations, without space-specific training.

5. Conclusions

This study provides the first investigation, to our knowledge, of the application of policy-gradient RL assisted by imitation learning to the integrated control of dynamic passive cooling and heating systems in buildings. Together, the results above support four central conclusions regarding this new method and the policies that result:

1. Imitation learning effectively captures expert knowledge of shading and natural ventilation control. Imitation learning, used to develop an appropriately structured neural network, generated a policy that mimicked the actions of a rule-based expert with high fidelity: shading actions were reproduced with over 95% accuracy, and natural ventilation actions were reproduced with over 80% accuracy. Additionally, the imitation network performed well across climate types, reducing early-summer space conditioning (cooling + heating) loads by over 45% in the humid subtropics, represented by cities of the Southeast U.S., and by over 70% in semi-arid, continental, and Mediterranean climates, represented by U.S. cities of the West and Midwest, North and Northeast, and Pacific Coast, respectively. These results showed that imitation learning has excellent potential to initialize RL policies with prior knowledge, supporting efforts to reduce total training time and to minimize undesirable actions during training.

2. RL training is improved by expert initialization. During subsequent RL training in contrasting climates, policies initialized with the imitation network converged in approximately 50% fewer iterations and achieved greater final performance than those initialized randomly. Whereas RL training with random initialization yielded control policies capable of reducing early-summer cooling loads by 30% or more in each of the four training climates, i.e. semi-arid Denver CO, continental Albany NY, humid subtropical Nashville TN, and Mediterranean Portland OR, these agents performed inconsistently among related climates in the corresponding regions. In some cases, particularly in the humid subtropical Southeast, their control actually increased these loads. RL agents initialized with the imitation network, in contrast, reduced cooling loads by 85% or more in each training city and by 50% or more in each related climate in the corresponding region. In direct comparisons, pre-trained RL agents reduced cooling loads to 30% or less of those reached by RL agents that lacked pretraining, in 22 of the 24 cities examined; in the two exceptions, pre-trained agents still out-performed the randomly initialized agents, but to lesser extents.

3. Climate-adapted RL policies are portable within broad climate types. RL training, following pre-training with the imitation network, generated observably different control patterns among the climates investigated: the policy adapted to continental climates, for example, added night insulation and lessened natural ventilation, maintaining warmer indoor conditions that avoided overcooling at night, while the humid subtropical policy increased morning and evening ventilation, maintaining cooler indoor air that minimized overheating during the day. This differentiation, as well as the improved performance among RL policies, illustrated the value of climate-adaptive training. Remarkably, these climate-adapted RL policies performed approximately as well, in numerous unfamiliar yet climatically related cities, as they did in the cities in which training occurred. Among the humid subtropical cities, the climate-adapted policy reduced early summer space conditioning loads by 55%–75%; in semi-arid, continental, and Mediterranean climates, climate-adapted policies reduced loads by 90%–99%. These trained policies were therefore portable across sizable geographic regions, each extending over 1000 km in its greatest dimension and encompassing substantial climatic variation. Climate specificity is a well-known limitation in the control of dynamic passive systems [e.g. 15,36]; these results show, for the first time, that effective control policies can be developed instead for broader climatic regions.

4. Climate-adapted RL policies are robust to building variations. Of equal significance, the climate-adapted RL policies performed virtually identically, within each climatic region, among dwellings that showed realistic variations in orientation, internal heat gain, window glazing materials, and air leakage rates. In Burlington VT, for example, the continental RL policy reduced cooling loads by over 99% in

each variation, while in Houston TX, the humid subtropical RL policy reduced cooling loads by approximately 50%–60% in each variation. The striking implication of this finding is that RL policies trained for a particular space type have unexpectedly low space-specificity, in notable contrast to rule-based controls, giving them the potential for far more general deployment.

Summary. The results above provide compelling evidence that passive heating and cooling systems can be well-controlled by regionally trained agents, without requiring extensive custom training in individual spaces. While the continuation of learning after an agent's deployment in a space is desirable as an ultimate goal, the evidence above indicates that substantial load reduction should be achievable immediately. This study therefore addresses four key challenges in passive cooling and heating control (Section 1.4), demonstrating a new solution for the design of passive cooling and heating control policies that require considerably reduced training; that avoid undesirable behavior during the learning stage; that provide excellent portability within regional climate types; and that maintain consistent performance throughout building variations.

Future work. Further refinements will be necessary to develop this work for application in occupied space. Annually complete expert policies and imitation networks must be developed, for example, and future reward structures should incorporate the value of daylighting and views as well as the relative energy and carbon profiles of the relevant heating and cooling systems. Responsiveness to weather forecasts will be advantageous, as well, to improve agents' abilities to accommodate unusually hot or cold weather. Modes of communication between agents and humans will also require thorough exploration to support the operation of manual systems, which prevail in residential spaces. The work presented here provides a rigorous foundation for such work: by documenting the performance, portability, and robustness of imitation-assisted reinforcement learning policies, it reveals that generalized control is possible, removing a central barrier to the realization of high-performance passive heating and cooling in residential buildings.

6. Code and product availability

All imitation and reinforcement learning codes, and EnergyPlus models, are available upon request.

CRediT authorship contribution statement

Bumsoo Park: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Alexandra R. Rempel:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Sandipan Mishra:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

The authors gratefully acknowledge the contributions of Joseph Bostick and Alan K.L. Lai to methodological development. This work was supported by the U.S. National Science Foundation's Environmental Sustainability Program through award CBET-1804218.

References

[1] International Energy Agency. Energy efficiency indicators database. 2022, URL https://www.iea.org/data-and-statistics/data-product/energy-efficiency-indicators-highlights.
[2] International Energy Agency. Global energy review: CO_2 emissions in 2021. 2021, URL https://iea.blob.core.windows.net/assets/c3086240-732b-4f6a-89d7-db01be018f5e/GlobalEnergyReviewCO2Emissionsin2021.pdf.
[3] Lucon O, Ürge-Vorsatz D, Zain Ahmed A, Akbari H, Bertoldi P, Cabeza L, et al. Climate change 2014: Mitigation of climate change: Contribution of working group III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press; 2014, p. 671–738.
[4] International Code Council. International energy conservation code (IECC), ch. 4 [RE]: Residential energy efficiency. 2018, Washington D.C., URL https://codes.iccsafe.org/content/iecc2018/chapter-4-re-residential-energy-efficiency.
[5] Passive House Institute US. PHIUS+ certification overview. 2021, URL https://www.phius.org/phius-certification-for-buildings-products/project-certification/overview.
[6] Garshasbi S, Haddad S, Paolini R, Santamouris M, Papangelis G, Dandou A, et al. Urban mitigation and building adaptation to minimize the future cooling energy needs. Sol Energy 2020;204:708–19.
[7] International Energy Agency. Renewable energy market update: May 2022. 2022, URL https://iea.blob.core.windows.net/assets/d6a7300d-7919-4136-b73a-3541c33f8bd7/RenewableEnergyMarketUpdate2022.pdf.
[8] Ellsworth-Krebs K. Implications of declining household sizes and expectations of home comfort for domestic energy demand. Nat Energy 2020;5(1):20–5.
[9] Gao J, Zhong X, Cai W, Ren H, Huo T, Wang X, et al. Dilution effect of the building area on energy intensity in urban residential buildings. Nature Commun 2019;10(1):1–9.
[10] International Energy Agency. Electricity market report. 2021, URL https://iea.blob.core.windows.net/assets/01e1e998-8611-45d7-acab-5564bc22575a/ElectricityMarketReportJuly2021.pdf.
[11] Waite M, Modi V. Electricity load implications of space heating decarbonization pathways. Joule 2020;4(2):376–94.
[12] Oropeza-Perez I, Østergaard PA. Energy saving potential of utilizing natural ventilation under warm conditions: A case study of Mexico. Appl Energy 2014;130:20–32.
[13] Tong Z, Chen Y, Malkawi A. Estimating natural ventilation potential for high-rise buildings considering boundary layer meteorology. Appl Energy 2017;193:276–86.
[14] Bastien D, Athienitis AK. A control algorithm for optimal energy performance of a solarium/greenhouse with combined interior and exterior motorized shading. Energy Procedia 2012;30:995–1005.
[15] Rempel AR, Danis J, Rempel AW, Fowler M, Mishra S. Improving the passive survivability of residential buildings during extreme heat events in the Pacific Northwest. Appl Energy 2022;321:119323.
[16] Rempel AR, Rempel AW, McComas SM, Duffey S, Enright C, Mishra S. Magnitude and distribution of the untapped solar space-heating resource in U.S. climates. Renew Sustain Energy Rev 2021;151:111599.
[17] Rempel AR, Lim S. Numerical optimization of integrated passive heating and cooling systems yields simple protocols for building energy decarbonization. Sci Technol Built Environ 2019;25:1226–36.
[18] Oropeza-Perez I, Østergaard PA. Active and passive cooling methods for dwellings: A review. Renew Sustain Energy Rev 2018;82:531–44.
[19] Van Moeseke G, Bruyère I, De Herde A. Impact of control rules on the efficiency of shading devices and free cooling for office buildings. Build Environ 2007;42(2):784–93.
[20] Tzempelikos A, Athienitis AK. The impact of shading design and control on building cooling and lighting demand. Sol Energy 2007;81(3):369–82.
[21] O'Donovan A, Murphy MD, O'Sullivan PD. Passive control strategies for cooling a non-residential nearly zero energy office: Simulated comfort resilience now and in the future. Energy Build 2021;231:110607.
[22] Rempel AR, Remington SJ. Optimization of passive cooling control thresholds with GenOpt and EnergyPlus. In: Proceedings of the symposium on simulation for architecture and urban design. 2015, p. 103–10.
[23] Tzempelikos A, Shen H. Comparative control strategies for roller shades with respect to daylighting and energy performance. Build Environ 2013;67:179–92.
[24] Wang Y, Kuckelkorn J, Liu Y. A state of art review on methodologies for control strategies in low energy buildings in the period from 2006 to 2016. Energy Build 2017;147:27–40.

[25] Yu L, Qin S, Zhang M, Shen C, Jiang T, Guan X. A review of deep reinforcement learning for smart building energy management. IEEE Internet Things J 2021;8(15):12046–63.
[26] Afram A, Janabi-Sharifi F. Theory and applications of HVAC control systems: A review of model predictive control (MPC). Build Environ 2014;72:343–55.
[27] Drgoňa J, Arroyo J, Figueroa IC, Blum D, Arendt K, Kim D, et al. All you need to know about model predictive control for buildings. Annu Rev Control 2020;50:190–232.
[28] Hu J, Karava P. Model predictive control strategies for buildings with mixed-mode cooling. Build Environ 2014;71:233–44.
[29] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT Press; 2018.
[30] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl Energy 2019;235:1072–89.
[31] Jia R, Jin M, Sun K, Hong T, Spanos C. Advanced building control via deep reinforcement learning. Energy Procedia 2019;158:6158–63.
[32] Osiński B, Jakubowski A, Zięcina P, Miłoś P, Galias C, Homoceanu S, Michalewski H. Simulation-based reinforcement learning for real-world autonomous driving. In: 2020 IEEE international conference on robotics and automation. IEEE; 2020, p. 6411–8.
[33] Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy 2020;269:115036.
[34] Sierla S, Ihasalo H, Vyatkin V. A review of reinforcement learning applications to control of heating, ventilation and air conditioning systems. Energies 2022;15(10):3526.
[35] Cheng Z, Zhao Q, Wang F, Jiang Y, Xia L, Ding J. Satisfaction based Q-learning for integrated lighting and blind control. Energy Build 2016;127:43–55.
[36] Chen Y, Norford LK, Samuelson HW, Malkawi A. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. Energy Build 2018;169:195–205.
[37] Ding X, Du W, Cerpa A. Octopus: Deep reinforcement learning for holistic smart building control. In: Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation. 2019, p. 326–35.
[38] Park B, Rempel AR, Lai AK, Chiaramonte J, Mishra S. Reinforcement learning for control of passive heating and cooling in buildings. IFAC-PapersOnLine (Special Issue: Modeling, Estimation and Control Conference, Austin TX) 2021;54(20):907–12.
[39] Han M, May R, Zhang X, Wang X, Pan S, Da Y, et al. A novel reinforcement learning method for improving occupant comfort via window opening and closing. Sustainable Cities Soc 2020;61:102247.
[40] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing atari with deep reinforcement learning. 2013, arXiv preprint arXiv:1312.5602.
[41] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[42] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 1992;8:229–56.
[43] Danis J, Mishra S, Rempel AR. Direct heat flux sensing for window shading control in passive cooling systems. Energy Build 2022;261:111950.
[44] Atzeri AM, Gasparella A, Cappelletti F, Tzempelikos A. Comfort and energy performance analysis of different glazing systems coupled with three shading control strategies. Sci Technol Built Environ 2018;24(5):545–58.
[45] da Silva PC, Leal V, Andersen M. Influence of shading control patterns on the energy assessment of office spaces. Energy Build 2012;50:35–48.
[46] Grynning S, Time B, Matusiak B. Solar shading control strategies in cold climates: Heating, cooling demand and daylight availability in office spaces. Sol Energy 2014;107:182–94.
[47] Roche L. Summertime performance of an automated lighting and blinds control system. Light Res Technol 2002;34(1):11–25.
[48] Firląg S, Yazdanian M, Curcija C, Kohler C, Vidanovic S, Hart R, et al. Control algorithms for dynamic windows for residential buildings. Energy Build 2015;109:157–73.
[49] Carletti C, Sciurpi F, Pierangioli L, Asdrubali F, Pisello AL, Bianchi F, Sambuco S, Guattari C. Thermal and lighting effects of an external venetian blind: Experimental analysis in a full scale test room. Build Environ 2016;106:45–56.
[50] Schulze T, Eicker U. Controlled natural ventilation for energy efficient buildings. Energy Build 2013;56:221–32.
[51] Schulze T, Gürlich D, Eicker U. Performance assessment of controlled natural ventilation for air quality control and passive cooling in existing and new office type buildings. Energy Build 2018;172:265–78.
[52] Liu M, Wittchen KB, Heiselberg PK. Control strategies for intelligent glazed façade and their influence on energy and comfort performance of office buildings in Denmark. Appl Energy 2015;145:43–51.
[53] Schaal S. Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences 1999;3(6):233–42.
[54] Abbeel P, Ng AY. Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the 21st international conference on machine learning. 2004, p. 1–8. http://dx.doi.org/10.1145/1015330.1015430.
[55] Bagnell J, Chestnutt J, Bradley D, Ratliff N. Boosting structured prediction for imitation learning. In: Schölkopf B, Platt J, Hoffman T, editors. Advances in neural information processing systems. 19, MIT Press; 2006, p. 1153–60, URL https://proceedings.neurips.cc/paper/2006/fil/fdbd31f2027f20378b1a80125fc862db-Paper.pdf.
[56] Attia A, Dayan S. Global overview of imitation learning. 2018, arXiv preprint arXiv:1801.06503.
[57] Ross S, Bagnell D. Efficient reductions for imitation learning. In: Proceedings of the 13th international conference on artificial intelligence and statistics, vol. 9. JMLR Workshop and Conference Proceedings; 2010, p. 661–8.
[58] Ross S, Gordon G, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the 14th international conference on artificial intelligence and statistics, vol. 15. JMLR Workshop and Conference Proceedings; 2011, p. 627–35.
[59] Ross S, Bagnell JA. Reinforcement and imitation learning via interactive no-regret learning. 2014, arXiv preprint arXiv:1406.5979.
[60] USDepartment of Energy. EnergyPlus version 9.2.0. 2019, URL https://energyplus.net.
[61] Big Ladder Software. Euclid v0.9.4.2: An open-source geometry editor for sketchup. 2017, Denver, CO, URL https://bigladdersoftware.com/projects/euclid/.
[62] Lawrence Berkeley National Laboratory. WINDOW 7.7Building Technologies & Urban Systems: Windows & Daylighting, Berkeley CA. 2019, URL https://windows.lbl.gov/tools/window/software-download.
[63] ANSI/ASHRAE. Standard 62.2: Ventilation and acceptable indoor air quality in residential buildings. 2019, American Society of Heating, Refrigerating Air-Conditioning Engineers, Atlanta, GA.
[64] US Department of Energy. EnergyPlus v.9.2.0 input output reference. 2019, Washington, D.C., URL https://energyplus.net.
[65] Lawrie L, Crawley D. Development of global typical meteorological years (TMYx). 2019, URL.
[66] US Department of Energy. EnergyPlus v.9.2.0 engineering reference. 2019, Washington, D.C., URL https://energyplus.net.
[67] Beck HE, Zimmermann NE, McVicar TR, Vergopolan N, Berg A, Wood EF. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci Data 2018;5(1):1–12.
[68] ANSI/ASHRAE. Standard 55-2020: Thermal environmental conditions for human occupancy. 2020, American National Standards Institute and the American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta GA.
[69] Dostal J. EnergyPlus co-simulation toolbox, v1.2.3.1.. 2021, URL www.mathworks.com/matlabcentral/fileexchange/69074-energyplus-co-simulation-toolbox.